View metadata, citation and similar papers at core.ac.uk

brought to you by CORE

provided by Research Online

University of Wollongong

Research Online

Centre for Statistical & Survey Methodology Working Paper Series

Faculty of Engineering and Information Sciences

2010

Mantel-Haenszel Estimators of Odds Ratios for Stratified Dependent Binomial Data

Thomas Suesse
University of Wollongong, tsuesse@uow.edu.au

Ivy Liu
University of Wollongong

Follow this and additional works at: https://ro.uow.edu.au/cssmwp

Recommended Citation

Suesse, Thomas and Liu, Ivy, Mantel-Haenszel Estimators of Odds Ratios for Stratified Dependent Binomial Data, Centre for Statistical and Survey Methodology, University of Wollongong, Working Paper 14-10, 2010, 36p.
https://ro.uow.edu.au/cssmwp/64

Research Online is the open access institutional repository for the University of Wollongong. For further information contact the UOW Library: research-pubs@uow.edu.au

Centre for Statistical and Survey Methodology

The University of Wollongong

Working Paper

14-10

Mantel-Haenszel Estimators of Odds Ratios for Stratified Dependent Binomial Data

Thomas Suesse, Ivy Liu

Copyright © 2008 by the Centre for Statistical & Survey Methodology, UOW. Work in progress, no part of this paper may be reproduced without permission from the Centre.

Centre for Statistical & Survey Methodology, University of Wollongong, Wollongong NSW 2522. Phone +61 2 4221 5435, Fax +61 2 4221 4845. Email: anica@uow.edu.au

Mantel-Haenszel Estimators of Odds Ratios for Stratified Dependent Binomial Data

Thomas Suesse[1] and Ivy Liu [2]
[1]: Centre for Statistical and Survey Methodology
University of Wollongong, Australia
[2] :School of Mathematics, Statistics and Operations Research
Victoria University of Wellington, New Zealand

January 13, 2011

Abstract

A standard approach to analyzing n binary matched pairs being usually represented in n 2×2 tables is to apply a subject-specific model; for the simplest situation it is the so-called Rasch Model. An alternative population-averaged approach is to apply a marginal model to the single 2×2 table formed by n subjects. For the situation of having an additional stratification variable with K levels forming K 2×2 tables, standard fitting approaches, such as generalized estimating equations and maximum likelihood, or alternatively the standard Mantel-Haenszel (MH) estimator can be applied. However, while all these standard approaches are consistent under a large stratum limiting model, they are not consistent under a sparse-data limiting model. In this paper, we propose a new MH estimator along with a variance estimator that are both dually consistent; consistent under large stratum and under sparse data limiting situations. In a simulation study the properties of the proposed estimators are confirmed and the estimator is compared with standard marginal methods, and also with subject-specific estimators. The simulation study also considers the case when the homogeneity assumption of the odds ratios does not hold and the asymptotic limit of the proposed MH estimator under this situation is derived. The results show that the proposed MH estimator is generally better than the standard estimator, and the same can be said about the associated Wald-type confidence intervals.

Keywords: bootstrap method, dual consistency, generalized estimating equations, Mantel-Haenszel estimator, odds ratio.

1 Introduction

Binary matched pairs data $\{\mathbf{Y}_i = (Y_{i1}, Y_{i2}),\ i = 1, \ldots, n\}$ frequently occur in medical trials and biological or social sciences. One approach to compare the probability of success for two binary responses $Y_{i1}, Y_{i2} \in \{0, 1\}$, here also called items, uses subject-specific models, and another uses marginal models. A popular subject-specific model is the Rasch model [1]

$$\text{logit}\left(\Pr(Y_{ix} = 1)\right) = \alpha_i + \beta^{sub} \cdot \mathbb{1}_{\{x=2\}}, \quad i = 1, \ldots, n \tag{1}$$

which assumes that Y_{i1} and Y_{i2} are independent, conditionally on α_i; this is called the local independence assumption. Here $\mathbb{1}_{\{cond\}}$ is the indicator function, which is one if condition *cond* is true, otherwise it is zero. Let n^{00}, n^{01}, n^{10} and n^{11} denote the number of observations with outcomes $\mathbf{Y} = (0, 0)$, $(0, 1)$, $(1, 0)$ and $(1, 1)$ respectively, with $n = n^{00} + n^{01} + n^{10} + n^{11}$. Assume $\{\alpha_i\}$ are fixed effects; the conditional maximum likelihood (CML), which conditions on the sufficient statistics for $\{\alpha_i\}$, gives the consistent estimator $\hat{\beta}^{sub}_{CML} = \log(n^{10}/n^{01})$. Another popular estimator is the Mantel-Haenszel (MH) [2] estimator. When applied to each of the n 2×2 tables formed by the n matched pairs, it also gives $\hat{\beta}^{sub}_{MH} = \log(n^{10}/n^{01})$. In each of the 2×2 tables, the columns refer to two items and the rows refer to the binary responses (0 and 1). Alternatively, by treating $\{\alpha_i\}$ as random effects, model (1) has the form of a generalized linear mixed model (GLMM). Given that the sample log odds ratio $\log(n^{11}n^{00})/(n^{10}n^{01}) \geq 0$, the maximum likelihood (ML) method gives $\hat{\beta}^{sub}_{GLMM} = \log(n^{10}/n^{01})$, independently of the random effects distribution provided that consistency conditions are met [3]. All three estimators are equal: $\hat{\beta}^{sub}_{MH} = \hat{\beta}^{sub}_{CML} = \hat{\beta}^{sub}_{GLMM}$.

Independently of $\{\alpha_i\}$ being random or fixed effects, the subject-specific model implies that marginally there is a non-negative correlation between Y_{i1} and Y_{i2} unless $\alpha_i = \alpha$. In the latter case, the Rasch model becomes

$$\text{logit}\left(\Pr(Y_{ix} = 1)\right) = \alpha + \beta^{ind} \cdot \mathbb{1}_{x=2} \tag{2}$$

and the ML method yields $\hat{\beta}^{ind}_{ML} = \log\{(n^{1+}n^{+0})/(n^{0+}n^{+1})\}$, where the notation "+" denotes the sum over that index. The MH estimator also gives $\hat{\beta}^{ind}_{MH} = \log\{(n^{1+}n^{+0})/(n^{0+}n^{+1})\}$ when applied to the single 2×2 table formed by the n subjects. When the Rasch model has the form of model (2) then Y_{i1} and Y_{i2} are marginally independent.

A marginal model has the same form as model (2)

$$\text{logit}\left(\Pr(Y_{ix} = 1)\right) = \alpha + \beta^{pop} \cdot \mathbb{1}_{x=2}, \tag{3}$$

but uses two common fitting procedures: ML or generalized estimating equations (GEE) [4] approaches, to take the dependency between Y_{i1} and Y_{i2} into account. Model (3) has population-averaged effect β^{pop}. Both approaches yield consistent estimates. For a good

summary see Agresti [5, Chapters 6, 10, 11, 12, 13]. Both β^{ind} and β^{pop} have marginal interpretations, population-averaged effects over all n subjects.

This paper considers the case of a confounding variable with K levels using marginal models. For example a clinical trial could be conducted in several, say K, hospitals. To be more general, we consider multiple ($c \geq 2$) binary observations per subject. The observations for the ith subject in stratum k are $\mathbf{Y}_{ik} = (Y_{i1k}, Y_{i2k}, \ldots, Y_{ick})$, where $i = 1, \ldots, n_k$, $k = 1, \ldots, K$ with $Y_{ixk} \in \{0, 1\}$. These binary observations form c dependent binomial counts $(Y_{1k}, Y_{2k}, \ldots, Y_{ck})$, where $Y_{xk} = \sum_{i=1}^{n_k} Y_{ixk}$ for $x = 1, \ldots, c$.

Liang [6] considered a marginal approach using the standard Mantel-Haenszel (MH) estimator for dependent binomial responses, but for a different dependence situation. He assumes the binary responses $Y_{1xk}, \ldots, Y_{n_k xk}$ that form the binomial responses Y_{xk} for item x are dependent, but items Y_{ixk}, Y_{iyk} for $x \neq y$ are still assumed to be independent. For matched pairs the opposite is true, in which items are dependent and binary observations Y_{ixk} forming the binomial responses Y_{xk} are assumed to be independent.

An example for $c = 2$ responses per subject is given in Table 1 and it shows both patients' self evaluation and investigators' evaluation on the patients' change in condition who suffer from asthma, at the end of the study, conducted by Merck Research Laboratories. The data were stratified by $K = 21$ clinical centers with 4 treatments. Let Y_{i1k} be the patient's self evaluation and Y_{i2k} be the evaluation from an investigator with values 0 (no change) and 1 (better) for the ith patient in the kth center. The table shows $n_{12|k}^{00}$, $n_{12|k}^{01}$, $n_{12|k}^{10}$ and $n_{12|k}^{11}$ the number of patients with outcomes $(Y_{i1k}, Y_{i2k}) = (0,0)$, $(0,1)$, $(1,0)$ and $(1,1)$, respectively for the kth clinical center for each treatment. One question researchers might ask is whether or not investigators are more positive towards the patients' improvement than the patients are themselves, controlling on clinical centers.

For this type of data, we can use either the subject-specific or the marginal approach [7, 8], depending on the research question and the interpretation needed. This paper focuses on the marginal approach with a possible stratifying variable, but not stratified on the finest level as subject. Such a marginal approach might be more appealing to clinicians, because in a clinical trial the investigator may be more interested in a population averaged effect, an effect averaged over the patients. The stratification could lead to a very sparse data set in the sense that within each stratum there are very few subjects. For instance, a clinical study might use many clinics because of the time it takes each clinic to recruit a sufficient number of patients. After stratifying the data according to the possible confounding variable (here clinics), the data become sparse.

Let $\pi_{x|k} = \Pr(Y_{xk} = 1)$ denote the (marginal) probability of a positive response and $\bar{\pi}_{x|k}$ denote the (marginal) probability of a negative response for item $x = 1, \ldots, c$ when a subject lies in stratum $k = 1, \ldots, K$. Now we focus on the more general marginal model with c binary responses

$$\text{logit}(\pi_{x|k}) = \alpha_k + \beta_x^{pop}, \quad x = 1, \ldots, c, \ k = 1, \ldots, K. \tag{4}$$

Regardless of the constraint for $\{\beta_x^{pop}\}$, it follows that the log odds ratio ($\log \Psi_{xy|k} = \log \frac{\pi_{x|k}\bar{\pi}_{y|k}}{\bar{\pi}_{x|k}\pi_{y|k}}$) equals $\beta_x^{pop} - \beta_y^{pop} =: \beta_{xy}^{pop}$ for all strata k. This is independent of k and implies a common odds ratio $\Psi_{xy} = \Psi_{xy|k}$ for all strata. For instance, we might assume that the effect Ψ_{12} describing the difference between the patients' self evaluation and the investigators' evaluation on the patients' change in condition is equal for all clinics.

The estimate of β_{xy}^{pop} can be obtained using the MH, ML and GEE methods. Under the standard assumption (e.g. [9, 10, 11]) that the underlying joint distribution $(Y_{i1k}, Y_{i2k}, \ldots, Y_{ick})$ follows a multinomial distribution, implying that the pairwise binary responses (Y_{ixk}, Y_{iyk}) with $x \neq y$ are also multinomials, we show that the standard MH estimator is not consistent under a sparse-data limiting situation (called limiting model II) for which $K \to \infty$, while all n_k (the total number of subjects in stratum k with $n_k = n_k^{00} + n_k^{01} + n_k^{10} + n_k^{11}$) remain bounded. All of the standard methods (MH, ML and GEE) are only consistent under the large-stratum limiting model (called limiting model I) (K is bounded, whereas $n_k \to \infty$). This paper proposes a new MH estimator and its variance estimator that are dually consistent, i.e. consistent under both limiting models I and II. We expect the new MH estimator to perform well under a sparse-data situation, e.g. as for clinical trial data comprising multiple centers, and none of the other marginal fitting methods to perform well in this situation.

The new MH estimator and its variance estimator are introduced in Section 2. In Section 3, we consider the ML and GEE methods for the marginal approach. Section 3 also illustrates the subject-specific approach through the CML and GLMM methods. When the items are positively correlated, the subject-specific effect is different from the population-averaged effect. In addition, we discuss the situation that the population-averaged effect and the subject-specific effect take a similar value. Section 4 discusses a meta-analysis situation not assuming a common odds ratio, where $\log \Psi_{xy|k}$ follows a normal or uniform distribution with mean $\log \Psi_{xy}$, the main effect. We show that the MH estimator is not consistent for $\log \Psi_{xy}$ under the sparse-data limiting model, but its limit converges to $\delta \log \Psi_{xy}$ with some constant $\delta \leq 1$. A simulation study is presented in Section 5, where all methods are compared in terms of mean square errors and coverage of a 95% confidence interval for the true parameter under various cases comprising negative and positive correlations as well as independence. Section 6 shows the results using different methods for the clinical trial example. The paper finishes with a general discussion, given in Section 7, that also provides some further useful applications for the new MH estimator when a population-averaged interpretation is sought.

2 Mantel-Haenszel (MH) Method

In this section we focus on the MH [2] estimators for the population-averaged effects $\{\beta_{xy}^{pop}, x \neq y = 1, \ldots, c\}$ of model (4). The notations for the MH estimators follow closely

the notations used by Greenland [12]. Let $n_{x|k}$ denote the number of positive responses for item $x = 1, \ldots, c$ and stratum $k = 1, \ldots, K$, and let $\bar{n}_{x|k}$ denote the number of negative responses. The standard MH estimator has the form

$$\hat{\Psi}_{xy} = C_{xy}/C_{yx} \tag{5}$$

with $C_{xy} = \sum_{k=1}^{K} c_{xy|k}$ and $c_{xy|k} = n_{x|k}\bar{n}_{y|k}/n_k$. Define $L_{xy} := \log \hat{\Psi}_{xy}$ as the estimator for the log odds ratio $\log \Psi_{xy}$. The MH estimator for β_{xy}^{pop} is L_{xy}.

Since $\mathbf{Y}_{ik} = (Y_{i1k}, Y_{i2k}, \ldots, Y_{ick})$ is a binary vector of length c with $i = 1, \ldots, n_k$ and $k = 1, \ldots, K$, there are 2^c possible outcomes per stratum $k = 1, \ldots, K$. Denote the cell counts for all 2^c possible outcomes by $\{n^j, j = 1, \ldots, 2^c\}$. We assume the underlying distribution of $\{n^j : j = 1, \ldots, 2^c\}$ is multinomial with parameters n_k and $\{\pi^j : j = 1, \ldots, 2^c\}$. The marginal probabilities $\{\pi_{x|k} : x = 1, \ldots, c\}$ can be computed from the joint probabilities $\{\pi_k^j : j = 1, \ldots, 2^c\}$ by a linear transformation.

Table 1 shows the complete clinical trial data with all $2^2 = 4$ ($c = 2$) possible outcomes for each clinic. Similarly, let the pairwise observations $(n_{xy|k}^{00}, n_{xy|k}^{10}, n_{xy|k}^{01}, n_{xy|k}^{11})$ follow a multinomial distribution with parameters n_k (stratum total) and $(\pi_{xy|k}^{00}, \pi_{xy|k}^{10}, \pi_{xy|k}^{01}, \pi_{xy|k}^{11})$ (the pairwise probabilities), where $\pi_{xy|k}^{st}$ is the probability of observing the pairwise outcome (s, t) for items x and y in the kth stratum, with $s, t \in \{0, 1\}$. The pairwise observations can be obtained from the joint observations. For the clinical trial example both the pairwise and full joint distribution are the same because $c = 2$, but often multiple items are recorded and then this distinction must be made. For the clinical trial example, $\pi_{12|k}^{11}$ is the probability that both the patient himself/herself and the investigator evaluated the change in condition as "better" in stratum k.

Greenland [12] considered two sampling models for each $2 \times c$ table: a) two rows of multinomial observations, and b) c independent binomials per stratum. He showed that the MH estimator (5) is dually consistent under both sampling models and also proposed dually consistent (co)variance estimators. The case of dependent binomial data, as in Table 1, can be viewed as an extension of case b).

We can show that the standard MH estimator $\hat{\Psi}_{xy}$ is not dually consistent under the dependent binomial case. From $\hat{\Psi}_{xy} - \Psi_{xy} = \Omega_{xy}/C_{yx}$ with $\Omega_{xy} = \sum_k \omega_{xy|k}$ and $\omega_{xy|k} = c_{xy|k} - \Psi_{xy} c_{yx|k}$, we obtain

$$\mathbb{E}\Omega_{xy} = \mathbb{E}(C_{xy} - \Psi_{xy} C_{yx}) = (1 - \Psi_{xy}) \sum_k (\pi_{x|k}\pi_{y|k} - \pi_{xy|k}^{11})$$

using $\mathbb{E}n_{x|k}n_{y|k} = n_k[n_k'\pi_{x|k}\pi_{y|k} + \pi_{xy|k}^{11}]$ with $n_k' := n_k - 1$. In order for the estimator to be consistent under limiting model II, we need $\mathbb{E}\Omega_{xy} = 0$. This happens only for the case $\Psi_{xy} = 1$ or $\pi_{xy|k}^{11} = \pi_{x|k}\pi_{y|k}$. The latter is true, when items x and y are (conditionally, given k) independent as in sampling model b). Appendix A shows the argument in detail and the inconsistency is confirmed by the simulation study that follows.

Under limiting model I, the MH estimator $\hat{\Psi}_{xy}$ is still consistent. This consistency also holds for both Greenland's co- and variance estimators. Since the MH estimator is not dually consistent anymore, it is likely to perform poorly under a sparse-data situation. The simulation study in Section 5 shows the performance of the standard MH estimator under both limiting models.

For the dependent binomial case – an extension of case b), we propose the following new estimator for the common odds ratio Ψ_{xy}

$$\tilde{\Psi}_{xy} = \frac{\tilde{C}_{xy}}{\tilde{C}_{yx}}$$

with $\tilde{C}_{xy} = \sum_k \tilde{c}_{xy|k}$ and $\tilde{c}_{xy|k} = (n_{x|k}\bar{n}_{y|k} - n^{10}_{xy|k})/n_k$; where by definition $n^{10}_{xy|k} = n^{01}_{yx|k}$. Note, $\tilde{c}_{xy|k}$ differs from $c_{xy|k}$ only by the extra term $n^{10}_{xy|k}$. Also, define $\tilde{L}_{xy}$ as the new estimator for $\log \Psi_{xy}$, i.e., $\tilde{L}_{xy} = \log \tilde{\Psi}_{xy}$. Since $\mathbb{E}\tilde{\Omega}_{xy} = 0$, where $\tilde{\Omega}_{xy} = \sum_k \omega_{xy|k}$ with $\omega_{xy|k} = \tilde{c}_{xy} - \Psi_{xy}\tilde{c}_{yx}$, it follows that $\tilde{\Psi}_{xy}$ is consistent under limiting model II. Under limiting model I, the additional terms (compared to $\hat{\Psi}_{xy}$) converge to zero. We conclude that the new estimator is indeed dually consistent.

Furthermore, we propose the following dually consistent variance estimator of $\tilde{L}_{xy}$

$$\begin{aligned}\tilde{U}_{xyy} := \widehat{\mathrm{Var}}(\tilde{L}_{xy}) = &\frac{\sum_k \frac{1}{n_k^2}((n^{10}_{xy|k})^2 - n^{10}_{xy|k})}{\tilde{C}^2_{xy}} + \frac{\sum_k \frac{1}{n_k^2}((n^{01}_{xy|k})^2 - n^{01}_{xy|k})}{\tilde{C}^2_{yx}} \\ &+ \frac{\sum_k \frac{n''_k n'_k + 2n'_k - 1}{n_k^2}(n^{10}_{xy|k} + n^{01}_{xy|k}) + \frac{2}{n_k^2} n^{10}_{xy|k} n^{01}_{xy|k} - \frac{n''_k}{n_k^2}(n^{10}_{xy|k} - n^{01}_{xy|k})^2}{\tilde{C}_{xy}\tilde{C}_{yx}}\end{aligned} \tag{6}$$

where $n'_k = n_k - 1$ and $n''_k = n_k - 2$. Appendix A also gives the detail of the proof for the dual consistency of $\tilde{\Psi}_{xy}$ and its variance estimator. The following theorem summarizes the findings.

Theorem 1 *The new MH estimator $\tilde{L}_{xy}$ and its new variance estimator $\tilde{U}_{xyy}$ are dually consistent for $\log \Psi_{xy}$ in marginal model (4). The old MH estimators L_{xy} and (co)variance estimators are consistent under limiting model I, and only consistent under model II for $\Psi = 1$ or conditional independence.*

Unfortunately, it does not seem feasible to provide new covariance estimators, because of the complexity involved in computing the covariance of $\tilde{L}_{xy}$ and $\tilde{L}_{xz}$ (or $\tilde{L}_{xy}$ and $\tilde{L}_{wz}$). To compute these covariances, we need to calculate higher moments based on the joint distribution with 3 (or 4) items comprising $2^3 = 8$ (or $2^4 = 16$) joint probabilities. This is

far more complex than computing the higher moments for the variance of $\tilde{L}_{xy}$ involving only $2^2 = 4$ (pairwise) probabilities.

As an alternative estimate for the co- and variance of the new MH estimator, the nonparametric bootstrap method [13] can be used, which randomly selects subjects with replacement from the original data forming a new artificial sample from which a new estimate $\hat{\Psi}$ is computed. Repeating this say $B = 1,000$ times creates a new sample of $\hat{\Psi}_1, \ldots, \hat{\Psi}_B$, from which the sample variance can be calculated, the so-called bootstrap estimate of variance.

The variance estimators (MH type or Bootstrap variance) are used to construct Wald-type confidence intervals (CI) to test for significance of the parameter Ψ_{xy}. There are other more sophisticated bootstrap CI methods, such as the percentile method, the bootstrap t-method (studentized pivotal) and the bias corrected accelerated (BCa) method, which are all implemented in the function boot.ci of `R`-package boot [14, 15]. Carpenter and Bithell [16] provide a good practical guide for bootstrap confidence intervals. For our situation they recommend either the BCa method or the variance stabilizing bootstrap-t method – an extension of the bootstrap-t method not provided by any `R`-package. This extension is preferred over the bootstrap-t method when the estimator and its variance estimator are not (approximately) independent. Figure 1 shows a typical situation for both the standard and the new MH estimators versus their variance estimators for $B = 1000$ bootstrap samples. Apparently the standard MH estimator has a slight linear relationship with its variance estimator, in contrast to the new MH estimator which seems almost independent of its variance estimator. This suggests that bootstrap CI methods provided by boot.ci are sufficient for our purposes. The simulation study that follows will also investigate the coverage of the various confidence intervals.

3 Alternative Marginal and Subject-Specific Approaches

3.1 Maximum Likelihood Method for the Marginal Approach

One approach to fit model (4) maximizes the multinomial likelihood for the K stratified 2^c tables while treating the model formula (4) as a set of constraint equations. Model (4) is a generalized log-linear model (GLLM) of the form

$$\mathbf{C} \log \mathbf{A}\boldsymbol{\pi} = \mathbf{X}\boldsymbol{\beta} \tag{7}$$

where $\boldsymbol{\pi}$ is a vector containing all $K \cdot 2^c$ joint probabilities; $\mathbf{C}$ and $\mathbf{A}$ are matrices; $\mathbf{X}$ is the design matrix and $\boldsymbol{\beta}$ is the vector of model parameters. Haber [17] and Lang and Agresti [9] presented numerical algorithms for maximizing multinomial likelihoods subject to constraints, i.e. model (7).

Many of the popular statistical packages do not have procedures available for ML fitting of such marginal models. An `R` function (mph.Rcode.R) for the algorithm may be obtained

from Prof J. B. Lang of the Statistics Department, University of Iowa http://www.stat.uiowa.edu/~jblang/. This R function can fit a multinomial Poisson homogeneous model, which is a wider class of model containing GLLM [11]. Bergsma et al. [18] proposed another fitting algorithm directly built on the work of Lang and Agresti [9] and Lang [10]. They provided an R package called "cmm" for fitting such models. Their program is a modification of the Lang-Agresti algorithm.

The ML estimator is not consistent for the sparse-data limiting model. The discussion is as follows. The log-likelihood kernel ll has the following form

$$ll = \sum_{k=1}^{K} \sum_{j=1}^{2^c} n^j \log \pi_k^j,$$

where j is the index referring to one of the 2^c outcomes. For simplification, we consider the case $c = 2$ with items x and y. Now ll depends on the $2^c = 4$ joint observations $n_{xy|k}^{00}$, $n_{xy|k}^{01}$, $n_{xy|k}^{10}$, $n_{xy|k}^{11}$ and multinomial probabilities $\pi_{xy|k}^{00}$, $\pi_{xy|k}^{01}$, $\pi_{xy|k}^{10}$, $\pi_{xy|k}^{11}$. The marginal probabilities are computed by $\pi_{x|k} = \pi_{xy|k}^{10} + \pi_{xy|k}^{11}$ and $\pi_{y|k} = \pi_{xy|k}^{01} + \pi_{xy|k}^{11}$ and a similar linear transformation applies to the observations. We have

$$\begin{aligned}
ll &= \sum_{k=1}^{K} \{n_{xy|k}^{00} \log \pi_{xy|k}^{00} + n_{xy|k}^{01} \log \pi_{xy|k}^{01} + n_{xy|k}^{10} \log \pi_{xy|k}^{10} + n_{xy|k}^{11} \log \pi_{xy|k}^{11}\} \\
&= \sum_{k=1}^{K} n_{xy|k}^{00} \log(1 - \pi_{xy|k}^{01} - \pi_{xy|k}^{10} - \pi_{xy|k}^{11}) + n_{xy|k}^{01} \log \pi_{xy|k}^{01} + n_{xy|k}^{10} \log \pi_{xy|k}^{10} + n_{xy|k}^{11} \log \pi_{xy|k}^{11} \\
&= \sum_{k=1}^{K} n_{xy|k}^{00} \log \left(1 - \text{expit}(\alpha_k + \beta_x^{pop}) - \text{expit}(\alpha_k + \beta_y^{pop}) + \pi_{xy|k}^{11}\right) \\
&+ n_{xy|k}^{01} \log \left(\text{expit}(\alpha_k + \beta_x^{pop}) - \pi_{xy|k}^{11}\right) + n_{xy|k}^{10} \log \left(\text{expit}(\alpha_k + \beta_y^{pop}) - \pi_{xy|k}^{11}\right) + n_{xy|k}^{11} \log \pi_{xy|k}^{11}
\end{aligned}$$

using $\pi_{x|k} = \text{expit}(\alpha_k + \beta_x^{pop}) = \frac{\exp(\alpha_k + \beta_x^{pop})}{1 + \exp(\alpha_k + \beta_x^{pop})}$ given by (4). Maximizing ll subject to model (4) means that we do not only obtain estimates for $\{\alpha_k : k = 1, \ldots, K\}$, β_x^{pop} and β_y^{pop}, but also for the additional nuisance parameters $\{\pi_{xy|k}^{11} : k = 1 \ldots, K\}$. There are $2 \cdot K + 2$ parameters in total. For $c > 2$, ll depends on even more parameters. This number grows linearly with 2^c.

If the number of items c becomes too large then the fitting becomes infeasible. Another problem arises when the number of observations per stratum n_k is small. Then the number of parameters can become larger than the number of observations. If n_k is bounded and $K \to \infty$ (limiting model II), then the ML is not consistent, because the number of parameters goes to infinity as the number of observations $N = \sum_{k=1}^{K} n_k$ does [19]. The ML method yields consistent estimates only for limiting model I.

3.2 Generalized Estimating Equations for the Marginal Approach

The generalized estimating equations (GEE) method [4] is a multivariate extension of the quasi-likelihood method for which we do not need to specify the full joint distribution of the c items $(Y_{i1k}, \ldots, Y_{ick})$. It only needs the structure for how the variance depends on the mean and the correlation structure of the c items. For the latter, one can make a choice for the "working correlation", such as independence, exchangeable or unstructured. The "robust" standard error adjusts the standard error using a "sandwich" method to reflect what actually occurs for the data. As is the ML method, the GEE method is only consistent under limiting model I, not under limiting model II, because in this situation the number of parameters grows with K. GEE is a robust and relatively easy method to apply and even yields consistent estimates if the working correlation is misspecified. It is also widely implemented in almost all popular software packages.

3.3 Conditional Maximum Likelihood (CML) for the Subject-Specific Approach

The analog subject-specific model to marginal model (4) is

$$\text{logit}\left(\pi_{x|ik}\right) = \alpha_{ik} + \beta_x^{sub}, \text{ where } i = 1, \ldots, n_k, \ k = 1, \ldots, K, \ x = 1, \ldots, c. \tag{8}$$

The $\{\alpha_{ik}\}$ introduce a non-negative correlation marginally between items. This model implies a common log odds ratio $\log \Psi_{xy}^{sub} = \frac{\pi_{x|ik}\bar{\pi}_{y|ik}}{\bar{\pi}_{x|ik}\pi_{y|ik}} = \beta_x^{sub} - \beta_y^{sub} := \beta_{xy}^{sub}$, because this expression is independent of i and k. In general, $\beta_{xy}^{sub} \neq \beta_{xy}^{pop}$. The CML estimator is dually consistent under the assumption of local independence (given $\{\alpha_{ik}\}$, Y_{ixk} and Y_{iyk} are independent). The estimator is also efficient for this situation. The method can be applied to either fixed effects or random effects and does not depend on any assumption of the random effects distribution.

3.4 Generalized Linear Mixed Model (GLMM) for the Subject-Specific Approach

For model (8), assume $\alpha_{ik} = \alpha_i + \alpha_k$. We consider two GLMMs under the common odds ratio assumption. The first (MM1) assumes that both α_i and α_k are random effects, i.e. $\alpha_i \sim N(0, \sigma_{sub}^2)$ and $\alpha_k \sim N(0, \sigma_{strata}^2)$, whereas the second (MM2) assumes that the $\{\alpha_k, k = 1, \ldots, K\}$ are fixed effects. MM1 is dually consistent, but MM2 is not consistent under limiting model II, since for this situation the number of parameters grows with K.

For MM2, we have the following relationship between β_{xy}^{sub} and β_{xy}^{pop}:

$$\beta_{xy}^{pop} \approx \delta_{\sigma_{sub}^2} \cdot \beta_{xy}^{sub}$$

where $\delta_{\sigma^2} := (1+\gamma^2\sigma^2)^{-1/2} \approx (1+0.35\sigma^2)^{-1/2}$ with $\gamma = 16\sqrt{3}/(15\pi)$ [20]. If $\sigma^2_{sub} \approx 0$ ($\Rightarrow \delta_{\sigma^2_{sub}} \approx 1$), then the subject-specific effect β^{sub}_{xy} and the marginal effect β^{pop}_{xy} are approximately equal. Otherwise, if $\sigma^2_{sub} > 0$, then $\delta_{\sigma^2_{sub}} < 1$.

4 Heterogeneity of Odds Ratios

The assumption of a common odds ratio is crucial for the consistency of the various estimators. However this assumption is not always fulfilled. A marginal model that allows the heterogeneity of odds ratios across different strata has the form

$$\text{logit}\left(\pi_{x|k}\right) = \alpha_k + \beta_{xk} + \beta_x,\ k = 1,\ldots,K,\ x = 1,\ldots,c. \tag{9}$$

Regardless of the constraint, the odds ratio in the kth stratum is $\log \Psi_{xy|k} = \beta_{xyk} + \beta_{xy} = (\beta_{xk} - \beta_{yk}) + (\beta_x - \beta_y)$. We can assume $\{\beta_{xk}\}$ as random effects, which is appropriate under limiting model II. Possible assumptions of the distribution include $\beta_{xk} \sim N(0, \sigma^2_x)$ and $\beta_{xk} \sim U[-d, +d]$, denoted by model (N) and (U) respectively. Consider an extension of the GLMMs discussed in Subsection 3.4, the subject-specific models analog to model (9) are referred to as models MM3 and MM4, which correspond to MM1 and MM2 but with additional random effects $\{\beta_{xk}\}$. The log odds ratio $\log \Psi_{xy} = \beta_{xy} = \beta_x - \beta_y$ could be considered as the average treatment effect, as in a meta-analysis.

The MH estimator does not converge to the average treatment effect because the average treatment effect is not a linear function of $(\{\pi_{x|k},\ x = 1,\ldots,c\})$. A similar inconsistency argument on a non-linear transformation was also given by Cox [21] considering a different situation where a random sample follows Poisson distributions with different means. When the distribution is parameterized in terms of some non-linear function, the transformation of the sample mean is not consistent anymore. The following theorem shows the limit of the MH estimator under limiting model II when odds ratios vary across strata. To be more general, the theorem is applied to the case where the correlation between β_{xk} and β_{yk} equals ρ_{xy} ($|\rho_{xy}| \leq 1$), for $x \neq y \in \{1,\ldots,c\}$.

Theorem 2 *For limiting model II, the standard MH estimator L_{xy} (under conditional independence or $\Psi_{xy} = 1$) and the new MH estimator $\tilde{L}_{xy}$ both converge approximately to $\log \Psi_{xy}$ under model (U) for a small $d \leq 1$, and converge approximately to $\delta_{\sigma^2_{xy}} \log \Psi_{xy}$ under model (N). The term $\delta_{\sigma^2} = (1+\gamma^2\sigma^2)^{-1/2} \approx (1+0.35\cdot\sigma^2)^{-1/2}$ where $\gamma = 16\sqrt{3}/(15\pi)$ and $\sigma^2_{xy} = \{\sigma^2_x + \sigma^2_y - \rho_{xy}\sigma_x\sigma_y\}/2$.*

There is a limitation of Theorem 2 due to the numerical approximation for the term $\mathbb{E}_{\beta_{xk}}\text{expit}(\alpha_k + \beta_{xk} + \beta_x)$. The approximation performs well when $d \leq 1$ for model (U) and when $\sigma^2_x \leq 4$ for model (N). The detailed proof is given in Appendix B.

The theorem implies that the limit will either differ from $\log \Psi_{xy}$ approximately by factor $\delta_{\sigma^2_{xy}} \leq 1$ (the value 1 occurs when $\sigma^2_{xy} \approx 0$) for model (N) and by factor 1 for model (U) for a small $d \leq 1$. When $\log \Psi_{xy} = 0$, the limit goes to zero for both random effect models. If we are only interested in testing $\log \Psi_{xy} = 0$ against $\log \Psi_{xy} \neq 0$, it follows that the MH estimator can also be used for testing the same hypothesis. The same applies to any of the other marginal methods. The simulation study that follows investigates the performance of the various methods under this heterogeneity situation.

5 Simulation Study

We conduct a simulation study to investigate the performance of the proposed log odds ratio estimator $\tilde{L}_{xy}$ and its variance estimator $\tilde{U}_{xyy}$. The proposed estimator $\tilde{L}_{xy}$ is compared with both the standard MH estimator L_{xy} and with those of all the other introduced methods: GEE, ML, CML, MM1, MM2, MM3 and MM4. For the GEE method, we fit the model using an exchangeable correlation structure, for the GLMM method we use `R`-package `lme4` [22] that uses Laplace approximation to obtain ML estimates.

5.1 Homogeneity Assumption

To generate data, we first compute the marginal probabilities from given parameters α_k and β_x according to marginal model (4) under the common odds ratio assumption (homogeneity). Parameters α_k were generated from $N(0, 1)$. We let $\beta_x = \log \Psi_{xy}/2$ and $\beta_y = -\log \Psi_{xy}/2$. Thus, $\beta_{xy} = \log \Psi_{xy}$.

We consider pairwise associations among c items using the odds ratio $\Gamma_{xy|k}$

$$\Gamma_{xy|k} = \frac{\pi^{11}_{xy|k}\pi^{00}_{xy|k}}{\pi^{01}_{xy|k}\pi^{10}_{xy|k}}$$

From the marginal probabilities $\{\pi_{x|k}, x = 1, \ldots, c\}$ and the association parameters $\{\Gamma_{xy|k}, x \neq y = 1, \ldots, c\}$, we can compute the pairwise probabilities $\{\pi^{st}_{xy|k}, x \neq y = 1, \ldots, c;\ s, t = 0, 1\}$ [23]. For simplicity, we let $c = 2$ allowing a simple sampling from the pairwise distribution (for $c = 2$ this is equal to the joint distribution) and assume a constant association parameter $\Gamma = \Gamma_{12|k}$ for all strata $k = 1, \ldots, K$. We use $S = 1, 2, 3$ to represent the sampling method $\Gamma = 0.01, 1.00, 10.00$, respectively. The stratum sample sizes n_k are set constant with values 5, 20, 100. The odds ratio Ψ $(= \Psi_{12})$ takes values 1 and 4. The number of strata K varies from 5 to 100. In the simulation study, the scenarios range from ones for which limiting model I seems suitable to ones for which limiting model II seems appropriate. The number of simulations is 5000 for $K = 100$ and 10000 otherwise, adjusting for the computational burden of the particular configuration.

Table 2 summarizes the mean squared error (mse) relative to the best method, with the first column showing the values of K, n_k , Ψ and S. The lowest value is 1.00 which indicates that the particular method has the lowest mse among the introduced methods. For instance, a relative mse (rmse) of 1.20 indicates that the particular method's mse is 1.20 times larger than the mse of the best method for this configuration. There are two additional numbers in superscript and subscript. In superscript, the percentage of simulations for which the method did not converge is shown. If this number is "0", then the method converged for all generated data sets, whereas the percentage 0.0 indicates that the method did not converge for up to 4 of the $10,000$ ($4/10000 = 0.04 \approx 0.0$) simulations. The number in subscript shows the contribution of $bias^2$ to the mse in percentage. For example the value 20 indicates that $bias^2$ contributes to 20% of the method's mse. There is also a sign (either "+" or "−") attached to this number, showing whether the bias is positive or negative. The table also includes the rmse's of the bootstrap samples mean (from $B = 1,000$ replicates) for both L and $\tilde{L}$, denoted by L_{BT} and $\tilde{L}_{BT}$, respectively. The results of MM2 and MM4 are not shown, because they are generally worse than those of MM1 and MM3.

Table 3 shows the percentage of times (coverage) that the 95% confidence interval covers the true parameter $\log \Psi$. The subscript shows again the percentage of simulations for which the method did not converge. The table shows the results of the percentile method, because the results of the other bootstrap methods are generally worse and are not shown. The tables showing the rmse and the coverage use complex designs, but are useful in preserving space and in summarizing multiple information in a single table.

We also want to point out a few issues when reading the tables and making interpretations. Sometimes the relative mse might seem good or even be best for some method, e.g. the ML method in Table 2 has a relative mse of 1.0 in a configuration, however the percentage of non-convergence is 98%. For the same configuration the MH methods converged for all simulated data sets. The mse's are not comparable if their computation is based on very different sets of simulations, but are still shown for completeness. Ideally, we could compute the mse of all methods for which the ML method converged. However this is also problematic, since then the results shown in the tables might only refer to a very few simulations (sometimes to none at all) making any comparison meaningless.

When the marginal model holds, the estimator $\widehat{\log \Psi}$ obtained from any of the subject-specific models (CML and MM1-4) can be quite different from the true population-averaged parameter $\log \Psi$. In practice, the choice of models between marginal and subject-specific is based on the nature of the research interest depending on the data. All of the subject-specific models have a large rmse when the items are highly positively correlated ($S = 3$), especially for the random effect models. This is not surprising because the difference between the population-averaged and the subject-specific log-odds ratios depends on the variance of the subject random effects in a GLMM, see Subsection 3.4. When

the items are highly correlated the variance is larger and consequently the difference is bigger. Although the simulations are not based on a GLMM, intuition suggests that a similar conclusion should hold for a GLMM that approximates the true distribution in the Kullback-Leibler sense.

From Table 3, we can also find the impact of the negative association between responses ($S = 1$) on the random effect models. When the null hypothesis is true ($\Psi = 1$), the coverage of the true parameter is less than 95% based on the 95% confidence interval. It implies that we reject the null hypothesis more often than we should using the random effect models. When the association between items is independent or positive ($S = 2$ or $S = 3$), the coverage of the true parameter is closer to the 95%. Therefore unsurprisingly the random effect models do not perform well when a non-negative association between the pair of responses is questionable. In comparison, the proposed MH method ($\tilde{L}$) has coverage close to the 95% for all cases.

From the rmse results in Table 2, the new MH estimator is generally better for $S = 3$ than the standard MH estimator, except for independence of items ($S = 2$) and $\Psi = 1$, for which the standard MH estimator is still dually consistent. The standard MH estimator also seems better for $S = 1$: when there is a strong negative correlation between items present. When we look at the construction of $\tilde{\Psi}_{xy}$ and compare it to that of $\hat{\Psi}$, then we see that the numerator and denominator of $\tilde{\Psi}_{xy}$ contain an extra term n^{10}: $\tilde{c}_{xy|k} = c_{xy|k} - n_k^{10}/n_k$. A negative correlation of items ($\Gamma < 1$), implies $\pi^{11}\pi^{00} < \pi^{10}\pi^{01}$. Hence n_k^{10} and n_k^{01} will be also relatively large. This implies that $\tilde{c}_{xy|k}$ and $\tilde{c}_{yx|k}$ are relatively smaller (closer to zero) than $c_{xy|k}$ and $c_{yx|k}$. Under $\Gamma < 1$ there will be more strata with zero contribution to the MH estimator $\tilde{\Psi}$ than for $\hat{\Psi}$, which leads to more inaccuracy. The new MH estimator is better than the standard one when a positive correlation presents between items.

Table 3 shows a different picture. The new MH estimator is to be preferred over the standard MH estimator for all sampling situations $S = 1, 2, 3$. Even though the standard MH estimator has better performance according to the rmse table under $S = 1$, the percentage of times that the 95% confidence interval covers the true parameter is much smaller than 95%. It can be explained by the fact that the standard variance estimator has worse performance than the new one. In summary, when the common odds ratio assumption holds, we can see from Tables 2 and 3, that the method that is often good is the new MH method.

5.2 Heterogeneity Assumption

We also generate data under model (9), for which the common odds ratio assumption does not hold (heterogeneity). We require additional parameters σ_x^2, σ_y^2 and ρ_{xy}. We assume that $\boldsymbol{\beta}_k = (\beta_{xk}, \beta_{yk})^T$ follows a bivariate normal with $\sigma_x^2 = \sigma_y^2 = 0.25$ and $\rho = 0.0, 0.8$. For the choice of $\sigma_x^2 = \sigma_y^2 = 0.25$, the generated β_{xyk} will be in the interval

$\beta_{xy} \pm 1.96\sqrt{2 \cdot 1/4} \approx \beta_{xy} \pm 1.385$ when $\rho = 0.0$. Consequently, 95% of the generated $\Psi_{xy|k}$ will lie in the interval $(\Psi_{xy}/4,\ 4 \cdot \Psi_{xy})$. For example, when $\Psi = 4$ the interval is $(1, 16)$ and when $\Psi = 1$ the interval is $(1/4, 4)$. For $\rho = 0.8$ the intervals will be smaller. In practice, we would only apply the MH estimator for small/moderate deviations from the common odds ratio assumption. For such a choice, the simulation study covers the cases from a small to a large deviation.

Tables 4 and 5 show the relative mse and the coverage when the data was generated under the heterogeneity assumption for $\rho_{xy} = 0.8$. Since the results under $\rho_{xy} = 0$ are very similar to the case of $\rho_{xy} = 0.8$, we only report one case. The MH method is now relatively slightly worse compared to those in the homogeneity cases. We had shown that under a sparse data situation the MH estimator does not converge to the true log odds ratio Ψ for $\Psi \neq 1$, instead it is often underestimated. For the case of $\rho_{xy} = 0.8$, it converges to $0.975 \cdot \log \Psi$ based on Theorem 2. Table 4 confirms this, showing negative biases for sparse data situations. The table also shows positive biases for the generalized linear mixed models (GLMMs) indicating that subject-specific effects are larger than the population-averaged ones. However, under $S = 2$ (independent responses), the GLMMs are close to the marginal models.

From Table 5, the models MM3 and MM4 allowing heterogeneity among the odds ratios have better performance than the models MM1 and MM2. However, similar to the homogeneous case, the type I error is larger than the significance level when the non-negative association assumption does not hold ($S = 1$).

Table 6 shows the proportion of simulations in which the null hypothesis $\Psi = 1$ was rejected when the true Ψ equals 2. Under limiting model I, the ML and GEE methods give a higher power of the test compared to the MH methods in most of cases. When data become sparse ($K = 20$, $N_k = 5$), even though the ML and GEE are still better, most of the cases (from 35% to 97%) do not converge. The MH methods based on the percentile bootstrap are not stable compared to the MH methods based on the derived variance formulae. We do expect a bigger power for the GLMMs, because the type I error is larger than the significance level (5%), as given by Tables 3 and 5. Similarly, since the type I error for the standard MH estimator is also larger than the significance level when $S = 1$, its power becomes larger than the proposed MH method under this case. Based on all simulated tables, we conclude that the proposed MH method performs well in various situations.

6 Example: Merck Research Laboratories Data

For Table 1, we compare the odds of being positive towards the improvement from the patients' self evaluation with these from the investigators' evaluation for each treatment. Because the data are very sparse, it is not sensible to use MM2 (or MM4) by treating the

clinical centers as fixed effects. In addition, the ML method fails to converge using Lang's algorithm [11]. Table 7 only reports the estimates from the standard MH, the new MH, GEE, CML, MM1, and MM3 methods.

For treatment 1 (Placebo), the proposed MH log odds ratio is $\tilde{L} = 0.621$ with standard error (s.e.) 0.323 based on formula (6). The bootstrap standard error gives a similar result 0.346. The standard MH log odds ratio is $L = 0.850$ with standard error 0.444, which are different from the proposed MH method. We expect a worse performance for the standard MH estimator, because the standard MH estimator is not consistent under limiting model II. We conclude that the estimated odds of being positive towards the improvement from the patients' self evaluation are exp(0.621)= 1.86 times higher than those from the investigators' evaluation in the placebo treatment. The odds ratio is significantly different from 1 at the 10% level. Alternatively, the log odds ratio estimates from the subject-specific methods for CML and MM1 are 1.705 (s.e. 0.768) and 2.698 (s.e. 0.897). Both of them are significant at the 5% level.

The patients' self evaluation tends to be more optimistic when the treatments 1 (placebo) or 2 (low dose of active drug) were assigned. When the dose of active drug is high (treatment 3 or 4), the difference between the patients' self evaluation and the investigators' evaluation diminishes. For the subject-specific models, the pattern is the same as for the MH method, but the subject-specific estimate is larger than the marginal methods.

The last column in Table 7 presents the estimate for all treatments combined, that is, the total number of strata equals $21 \times 4 = 84$. It gives an estimate of the mean effect across all clinical centers and treatments, allowing for heterogeneity across strata. On average, the patients' self evaluation still tends to be more positive and the difference between the patients' self evaluation and the investigators' evaluation is significant at the 5% level. In summary, the estimated odds of being positive towards the improvement from the patients' self evaluation are exp(0.400)= 1.49 times higher than those from the investigators' evaluation.

7 Discussion

In this paper we propose a new MH estimator for stratified dependent binomial data. It has advantages over the GEE, ML, and standard MH estimators for sparse data, because the new MH estimator is consistent under limiting model II whereas the other three are not. When the data are not sparse, it performs as well as the GEE and ML estimators based on our simulation study. Unlike the standard MH variance estimator, the proposed dually consistent variance estimator also performs well, giving correct coverage of the true parameter. Another advantage of our method is that, e.g. in a multi-center study, the correlation between responses can be different for different centers or for treatments.

Unlike traditional approaches, the new MH method does not assume equal correlations in different strata.

The paper focuses on the marginal model, a population-averaged approach, but alternative subject-specific estimators can be obtained using the CML or random effect methods. In our view, both the population-averaged and subject-specific approaches have their justification. The choice between these two depends on the nature of the study. This paper points out the problems of using the subject-specific approach if a population-averaged effect is indeed the main interest.

Assume that the subject nuisance parameter follows a normal distribution. The magnitude of the difference between the population-averaged and the subject-specific effects depends on the variance of the nuisance parameters. The difference increases as the variance increases (i.e., the association between responses increases). The subject-specific effects are larger in absolute value (factor δ_{σ^2}), as discussed in Subsection 3.4. When the responses do not have positive associations, which violates the structure of the GLMM, the type I error becomes larger than the significance level. For the marginal model, the population-averaged effect is independent of the magnitude of the correlation between items. The negative association does not have any effect on the population-averaged estimator.

This paper also discusses the performance of the new MH estimator when the common odds ratio assumption does not hold. If the common odds ratio assumption is slightly violated it is still a useful tool in obtaining a summarizing effect. However, when the log odds ratios vary across strata, as in a meta-analysis situation, we show that the MH estimator converges to the mean log odds ratio with a factor δ (≤ 1). The factor δ is ≤ 1 under limiting model II. The value of δ depends on the variation of the log odds ratios across strata. The larger the variation, the worse the MH estimator is. For instance, in our simulation study the odds ratio varies from 1 to 16 when the true mean odds ratio equals 4. The MH estimator under-estimates the true mean odds ratio when the data are sparse, but in a small scale. For the above case, the MH estimator converges to $\exp(0.96 \cdot \log 4) = 3.78$ (instead of 4).

The new MH type estimator can be easily applied to any binary matched pairs data situation. For example let us focus on a multi-center cross-over study design with treatments A and B and a stratification variable with K levels referring to the K clinical centers. The standard approach considered by Gart [24] is based on the subject-specific approach. Instead we can also apply the newly proposed MH estimator based on a marginal model by first computing $\tilde{\Psi}_{AB}$ for the group that receives treatment A first, followed by treatment B. Similarly we can compute $\tilde{\Psi}_{BA}$ for the other group which received treatment B first. The treatment effect for the cross-over study is computed by $\frac{1}{4}(\log \tilde{\Psi}_{AB} + \log \tilde{\Psi}_{BA})$ and the order effect by $\frac{1}{4}(\log \tilde{\Psi}_{AB} - \log \tilde{\Psi}_{BA})$. Another application is a multi-center cohort study comparing two groups. The change for the control group between baseline

and follow-up is captured by the odds ratio Ψ_C and that of the treatment group by Ψ_T, assuming a common odds ratio for the K medical centers. The treatment effect can now be estimated by $\log(\tilde{\Psi}_T) - \log(\tilde{\Psi}_C)$, which accounts for the dependence between baseline and follow-up. This situation is similar to a matched-pairs situation for continuous data, for which usually the one-sample t-test is applied. Another application of the proposed estimators is a longitudinal study, in which the binary responses are collected over time.

This paper gives a simple method to compare binary matched pairs when the researcher is interested in a population-averaged interpretation and when the data were highly stratified by other factors. Another reason to opt for the population-averaged approach is that the marginal effect is independent of the magnitude of the association between items. In contrast the subject-specific effect increases with increasing positive correlation (factor $\delta_{\sigma^2_{sub}}$ with $\alpha_i \sim N(0, \sigma^2_{sub})$). Medical practitioners often compare treatment effects of several trials. When the treatment effects are based on mixed models with different magnitude of correlation within subjects, then differences in the treatment effects might be solely due to the difference in correlation, i.e. the difference in δ_{σ^2}, but not due to a different marginal effect (odds ratio referring to marginal probabilities).

Our proposed MH estimator can be generalized along the lines of Mickey and Elashoff [25] and Greenland [12]. The log odds ratio $\log \Psi_{xz}$ cannot only be estimated by L_{xz}, but also by $L_{xy} + L_{yz}$. This is because of the property $\Psi_{xz} = \Psi_{xy}\Psi_{yz}$. There does not exist a unique estimator, since generally $L_{xz} \neq L_{xy} + L_{yz}$. The generalized estimator introduced by Greenland [12]

$$\widehat{\log \Psi_{xy}} := \bar{L}_{xy} := (L_{x+} - L_{y+})/c.$$

is generally applicable to any estimator of $\log \Psi_{xy}$ and also applicable for stratified dependent binomial data. The generalized MH estimator has efficiency advantages over the standard MH estimator. Greenland [12] also proposed a dually consistent covariance estimator for the covariance between $\bar{L}_{xy}$ and $\bar{L}_{wz}$ under the sampling models a) and b). The formulae for such a covariance estimator for the proposed MH estimators is slightly different and is presented in Appendix C along with a proof.

Suesse [26] defined the new MH estimator $\tilde{\Psi}$ slightly differently. The term $\tilde{c}_{xy|k}$ in this paper uses denominator n_k, whereas Suesse [26] used denominator n'_k. Using n_k has the advantage that $\tilde{\Psi}$ is also automatically defined for $n_k = 1$, whereas using n'_k has the advantage that then $\mathbb{E}\tilde{c}_{xy|k} = n_k \pi_{x|k} \bar{\pi}_{y|k}$ (holds in general) is identical to $\mathbb{E}c_{xy|k} = n_k \pi_{x|k} \bar{\pi}_{y|k}$, i.e. standard and new MH estimator share the same property, provided the underlying assumptions are fulfilled.

Unfortunately we cannot use the new MH estimator when we condition on each subject i, because then $n_k = 1$ and the contribution of each subject to the MH estimator is zero ($\tilde{c}_{xy|k} = 0$), making the new MH estimator undefined in this case. This is in contrast to the standard MH estimator (assuming conditional independence) which is still defined in this instance.

The variance estimator $\tilde{U}_{xyy}$ can be improved in estimating the variance of $\tilde{L}$, even though the coverage of the Wald-type confidence interval is good. Future research aims at finding a more efficient variance estimator replacing $\tilde{U}_{xyy}$ as well as covariance estimators that go along with the asymptotic covariances yet to be derived. Then the generalized (co)variance estimators can also be constructed by using these more efficient estimators. It needs to be investigated whether deriving covariance estimators is feasible or not.

Following the discussion of effects of misclassification in matched-pair case-control studies given by Greenland [27], the proposed method along with any other considered method here, can be sensitive to bias from misclassification. Future research also aims to address the effect of misclassification on the newly proposed MH estimator.

Acknowledgements

We would like to thank Professors Alan Agresti and Sander Greenland for their helpful comments.

References

[1] Rasch G. On the general laws of and the meaning of measurement in psychology. *4th Berkeley Symposium on Mathematical Statistics and Probability*, vol. 4, 1961; 321–355.

[2] Mantel N, Haenszel W. Statistical aspects of the analysis of data from retrospective studies of disease. *Journal of the National Cancer Institute* 1959; **22**:719–748.

[3] Neuhaus JM, Kalbfleisch JD, Hauck WW. Conditions for consistent estimation in mixed-effects models for binary matched-pairs data. *Canadian Journal of Statistics-Revue Canadienne De Statistique* 1994; **22**(1):139–148.

[4] Liang KY, Zeger SL. Longitudinal data-analysis using generalized linear-models. *Biometrika* 1986; **73**(1):13–22.

[5] Agresti A. *Categorical Data Analysis*. 2nd edition edn., Wiley Series in Probability and Statistics, Wiley, 2002.

[6] Liang KY. Odds ratio inference with dependent data. *Biometrika* 1985; **72**(3):678–682.

[7] Agresti A, Liu IM. Modeling a categorical variable allowing arbitrarily many category choices. *Biometrics* 1999; **55**(3):936–943.

[8] Agresti A, Liu I. Strategies for modeling a categorical variable allowing multiple category choices. *Sociological Methods & Research* 2001; **29**(4):403–434.

[9] Lang JB, Agresti A. Simultaneously modelling joint and marginal distributions of multivariate categorical responses. *Journal of the American Statistical Association* 1994; **89**(426):625–632.

[10] Lang JB. Maximum likelihood methods for a generalized class of log-linear models. *Annals of Statistics* 1996; **24**(2):726–752.

[11] Lang JB. Multinomial-Poisson homogeneous models for contingency tables. *Annals of Statistics* 2004; **32**(1):340–383.

[12] Greenland S. Generalized Mantel-Haenszel estimators for K 2xJ tables. *Biometrics* 1989; **45**(1):183–191.

[13] Efron B, Tibshirani R. *An Introduction to the Bootstrap*. Chapman and Hall: New York, 1993.

[14] Davison AC, Hinkley DV. *Bootstrap Methods and Their Applications*. Cambridge University Press: Cambridge, 1997.

[15] Canty A, Ripley B. boot: Bootstrap R (S-Plus) Functions. R-package version 1.2-38 2009.

[16] Carpenter J, Bithell J. Bootstrap confidence intervals: when, which, what? A practical guide for medical statisticians. *Statistics in Medicine* 2000; **19**(9):1141–1164.

[17] Haber M. Maximum-likelihood methods for linear and log-linear models in categorical-data. *Computational Statistics & Data Analysis* 1985; **3**(1):1–10.

[18] Bergsma W, Croon M, Hagenaars J. *Marginal Models for Dependent, Clustered, and Longitudinal Categorical Data*. Springer: New York, 2009.

[19] Neyman J, Scott E. Consistent estimates based on partially consistent observations. *Econometrika* 1948; **16**(1):1–22.

[20] Zeger SL, Liang KY, Albert PS. Models for longitudinal data - a generalized estimating equation approach. *Biometrics* 1988; **44**(4):1049–1060.

[21] Cox DR. Some remarks on overdispersion. *Biometrika* 1983; **70**(1):269–274.

[22] Bates D, Maechler M. R-package: lme4: Linear mixed-effects models using S4 classes 2009.

[23] Bilder CR, Loughin TM. Testing for conditional multiple marginal independence. *Biometrics* 2002; **58**(1):200–208.

[24] Gart JJ. An exact test for comparing matched proportions in crossover designs. *Biometrika* 1969; **56**(1):75–80.

[25] Mickey RM, Elashoff RM. A generalization of the Mantel-Haenszel estimator of partial association for 2xJxK-tables. *Biometrics* 1985; **41**(3):623–635.

[26] Suesse T. Analysis and diagnostics of categorical variables with multiple outcomes. PhD Thesis, Victoria University 2008.

[27] Greenland S. The effect of misclassification in matched-pair case-control studies. *American Journal of Epidemiology* 1982; **116**(2):402–406.

Table 1: Patients' Self Evaluation and Investigators' Evaluation on the Patients Change in Condition: number of bivariate binary observations for improvement

Center	1	2	3	4	5	6	7
Treatments	1,2,3,4	1,2,3,4	1,2,3,4	1,2,3,4	1,2,3,4	1,2,3,4	1,2,3,4
n^{00}	4,1,2,0	2,1,2,2	4,1,1,3	1,2,2,1	2,1,2,0	0,0,0,0	1,3,1,2
n^{01}	0,0,1,1	1,0,0,0	0,0,0,0	0,0,0,0	0,0,0,1	0,0,0,0	0,0,0,1
n^{10}	0,2,0,1	0,1,1,1	1,3,0,0	1,2,2,1	0,0,0,1	2,0,0,1	2,1,2,0
n^{11}	0,1,1,1	0,0,0,0	1,0,0,2	0,2,2,0	1,1,1,3	0,2,1,1	0,0,0,2
Center	8	9	10	11	12	13	14
Treatments	1,2,3,4	1,2,3,4	1,2,3,4	1,2,3,4	1,2,3,4	1,2,3,4	1,2,3,4
n^{00}	2,1,0,2	1,3,2,0	1,2,0,1	2,1,1,1	0,0,0,1	4,3,1,2	1,2,1,3
n^{01}	0,0,1,1	0,0,0,0	0,0,0,0	0,0,0,1	0,0,0,1	0,0,0,1	0,0,0,0
n^{10}	0,0,0,0	0,0,0,0	0,0,0,0	1,0,1,0	0,0,0,1	1,0,0,0	0,1,0,1
n^{11}	0,1,2,0	0,0,1,1	1,2,1,1	0,1,1,0	2,1,2,0	0,1,1,0	0,2,1,1
Center	15	16	17	18	19	20	21
Treatments	1,2,3,4	1,2,3,4	1,2,3,4	1,2,3,4	1,2,3,4	1,2,3,4	1,2,3,4
n^{00}	1,5,3,1	2,2,2,0	4,0,3,2	4,1,0,1	1,0,3,3	1,2,3,1	0,3,1,1
n^{01}	0,0,0,0	0,1,0,0	0,0,0,0	1,0,0,0	0,0,0,0	0,0,0,1	0,1,0,0
n^{10}	2,0,0,0	0,1,0,0	0,0,0,2	0,1,0,0	1,0,0,0	0,1,0,0	0,0,0,0
n^{11}	1,0,2,1	2,1,3,4	1,1,2,0	0,2,1,1	1,3,1,2	1,0,1,0	1,0,2,2

Table 2: Relative Mean Squared Error (MSE) under common odds ratio assumption

K n_k Ψ S	L	$\tilde{L}$	L_{BT}	$\tilde{L}_{BT}$	GEE	ML	CML	MM1	MM3
5 20 1 1	$2.20^{0}_{0.0-}$	$2.32^{0}_{0.0-}$	$1.62^{0}_{0.0-}$	$1.67^{0}_{0.0-}$	2.44^{53}_{66+}	$2.32^{0}_{0.0-}$	$1.00^{0}_{0.0-}$	$2.19^{2.3}_{0.0-}$	$2.25^{4.7}_{0.0-}$
5 20 1 2	$1.40^{0}_{0.0+}$	$1.40^{0}_{0.0+}$	$1.02^{0}_{0.0+}$	$1.00^{0}_{0.0+}$	1.55^{53}_{67+}	$1.55^{0.1}_{0.0+}$	$1.45^{0}_{0.0+}$	$1.26^{1.1}_{0.0-}$	$1.47^{2.3}_{0.0+}$
5 20 1 3	$1.15^{0}_{0.0-}$	$1.10^{0}_{0.0-}$	$1.05^{0}_{0.0-}$	$1.00^{0}_{0.0-}$	1.28^{55}_{68+}	$1.61^{9.6}_{0.0-}$	$5.10^{0}_{0.0-}$	$1.21^{0}_{0.0-}$	$3.78^{3.8}_{0.0-}$
5 20 4 1	$1.10^{0}_{0.1-}$	$1.18^{0}_{0.5+}$	$1.00^{0}_{8.3-}$	$1.01^{0}_{4.2-}$	$1.18^{0}_{0.7+}$	$1.18^{0}_{0.7+}$	2.21^{0}_{78-}	$1.08^{1.4}_{0.6-}$	$1.13^{3.1}_{0.1+}$
5 20 4 2	$1.13^{0}_{0.7+}$	$1.13^{0}_{0.7+}$	$1.00^{0.5}_{20-}$	$1.01^{0.5}_{22-}$	$1.18^{9.6}_{2.3+}$	$1.23^{9.7}_{2.4+}$	$1.52^{0}_{1.9+}$	1.05^{0}_{26-}	$1.34^{2.1}_{4.0+}$
5 20 4 3	$1.05^{0}_{2.6+}$	$1.00^{0}_{0.4+}$	1.33^{0}_{56-}	1.49^{0}_{62-}	$1.09^{0.2}_{4.6+}$	$1.06^{0.4}_{2.9+}$	$20.1^{8.5}_{79+}$	$8.55^{6.7}_{72+}$	$9.55^{5.1}_{73+}$
20 5 1 1	$2.05^{0}_{0.0+}$	$2.60^{0}_{0.0+}$	$1.82^{0}_{0.0-}$	$2.19^{0}_{0.0-}$	2.85^{69}_{66+}	$2.81^{35}_{0.0+}$	$1.00^{0}_{0.0-}$	$2.20^{1.1}_{0.0-}$	$2.50^{2.8}_{0.0-}$
20 5 1 2	$1.82^{0}_{0.0+}$	$1.82^{0}_{0.0+}$	$1.11^{0}_{0.0+}$	$1.00^{0}_{0.0+}$	2.09^{97}_{70+}	$2.18^{97}_{0.1+}$	$1.89^{0}_{0.0+}$	$1.83^{1.6}_{0.0-}$	$1.91^{1.9}_{0.0+}$
20 5 1 3	$1.93^{0}_{0.0-}$	$1.57^{0}_{0.0-}$	$1.68^{0}_{0.0-}$	$1.34^{0}_{0.0-}$	2.35^{93}_{68+}	$1.00^{98}_{0.0-}$	$6.99^{0}_{0.0-}$	$4.84^{0.9}_{0.0-}$	$5.04^{2.1}_{0.0-}$
20 5 4 1	$1.03^{0}_{6.9-}$	$1.22^{0}_{0.7+}$	1.10^{0}_{34-}	1.00^{0}_{20-}	$1.33^{34}_{4.9+}$	$1.33^{34}_{5.0+}$	1.83^{0}_{71-}	$1.02^{3.3}_{6.0-}$	$1.13^{1.6}_{0.0-}$
20 5 4 2	$1.16^{0}_{0.4+}$	$1.18^{0}_{0.4+}$	$1.00^{0}_{8.4-}$	1.02^{0}_{14-}	1.68^{47}_{18+}	1.47^{67}_{20+}	$1.51^{0}_{1.2+}$	$1.33^{6.8}_{2.4+}$	$1.38^{1.7}_{3.6+}$
20 5 4 3	1.69^{0}_{17+}	$1.27^{0}_{0.4+}$	$1.00^{0}_{9.0-}$	1.39^{0}_{43-}	2.47^{76}_{38+}	$1.53^{84}_{3.9+}$	$24.9^{8.3}_{79+}$	$10.3^{7.9}_{76+}$	$12.0^{6.4}_{74+}$
5 100 1 1	$1.70^{0}_{0.0+}$	$1.71^{0}_{0.0+}$	$1.00^{0}_{0.0+}$	$1.00^{0}_{0.0+}$	1.78^{51}_{66+}	$1.72^{0}_{0.0+}$	$1.04^{0}_{0.0+}$	$1.68^{2.0}_{0.1+}$	$1.70^{2.3}_{0.0+}$
5 100 1 2	$1.88^{0}_{0.0+}$	$1.88^{0}_{0.0+}$	$1.01^{0}_{0.0+}$	$1.00^{0}_{0.0+}$	1.96^{51}_{65+}	$1.91^{0}_{0.0+}$	$1.89^{0}_{0.0+}$	$1.19^{4.3}_{0.0+}$	$1.92^{1.7}_{0.0+}$
5 100 1 3	$1.29^{0}_{0.0+}$	$1.28^{0}_{0.0+}$	$1.01^{0}_{0.0+}$	$1.00^{0}_{0.0+}$	1.35^{52}_{66+}	$1.33^{0}_{0.0+}$	$5.20^{0}_{0.0+}$	$4.03^{0}_{0.0+}$	$4.04^{0.8}_{0.0+}$
5 100 4 1	$1.01^{0}_{0.0-}$	$1.03^{0}_{0.2+}$	1.47^{0}_{47-}	1.43^{0}_{46-}	$1.03^{0}_{0.2+}$	$1.02^{0}_{0.2+}$	8.19^{0}_{95-}	$1.00^{2.8}_{0.0-}$	$1.02^{4.8}_{0.1+}$
5 100 4 2	$1.00^{0}_{0.2+}$	$1.00^{0}_{0.2+}$	2.48^{0}_{72-}	2.51^{0}_{72-}	$1.01^{0}_{0.7+}$	$1.03^{0}_{0.7+}$	$1.23^{0}_{0.4+}$	$1.06^{2.3}_{1.7+}$	$1.07^{1.9}_{1.5+}$
5 100 4 3	$1.01^{0}_{0.7+}$	$1.00^{0}_{0.2+}$	11.3^{0}_{97-}	11.6^{0}_{97-}	$1.02^{0}_{1.1+}$	$1.03^{0}_{1.2+}$	55.9^{0}_{90+}	$18.2^{0.1}_{88+}$	$18.2^{0.6}_{88+}$
100 5 1 1	$1.97^{0}_{0.0+}$	$2.44^{0}_{0.0+}$	$1.62^{0}_{0.0+}$	$1.86^{0}_{0.0+}$	NA^{100}	NA^{99}	$1.00^{0}_{0.0+}$	$2.14^{0.2}_{0.0+}$	$2.37^{0.2}_{0.0+}$
100 5 1 2	$1.37^{0}_{0.0-}$	$1.36^{0}_{0.0-}$	$1.06^{0}_{0.0-}$	$1.00^{0}_{0.0-}$	NA^{100}	NA^{100}	$1.38^{0}_{0.0-}$	$1.32^{3.4}_{0.1-}$	$1.38^{1.8}_{0.0-}$
100 5 1 3	$1.62^{0}_{0.0-}$	$1.31^{0}_{0.0-}$	$1.27^{0}_{0.0-}$	$1.00^{0}_{0.0-}$	NA^{100}	NA^{100}	$5.25^{0}_{0.0-}$	$3.68^{0.6}_{0.0-}$	$3.99^{1.6}_{0.0-}$
100 5 4 1	1.29^{0}_{38-}	$1.02^{0}_{0.1+}$	2.75^{0}_{78-}	1.92^{0}_{65-}	1.34^{96}_{14+}	1.31^{96}_{15+}	6.80^{0}_{94-}	$1.12^{3.6}_{26-}$	$1.00^{0.4}_{0.3-}$
100 5 4 2	$1.00^{0}_{0.0+}$	$1.01^{0}_{0.0+}$	1.68^{0}_{56-}	2.10^{0}_{67-}	NA^{100}	NA^{100}	$1.19^{0}_{0.1+}$	$1.00^{5.8}_{0.1+}$	$1.02^{1.6}_{0.6+}$
100 5 4 3	2.12^{0}_{48+}	$1.00^{0}_{0.0+}$	1.00^{0}_{21-}	2.84^{0}_{76-}	NA^{100}	NA^{100}	96.3^{0}_{93+}	$41.0^{0.1}_{93+}$	$41.1^{2.0}_{93+}$

Table 3: Coverage under common odds ratio assumption

K n_k Ψ S	L	$\tilde{L}$	L_{perc}	$\tilde{L}_{perc}$	GEE	ML	CML	MM1	MM2	MM3	MM4
5 20 1 1	89.6^{0}	95.7^{0}	94.7^{0}	94.8^{0}	95.1^{53}	95.2^{0}	95.4^{0}	$89.0^{2.3}$	$88.9^{0.9}$	$92.2^{4.7}$	$89.6^{3.5}$
5 20 1 2	94.8^{0}	95.1^{0}	94.5^{0}	94.6^{0}	94.4^{53}	$92.6^{0.1}$	95.2^{0}	$94.8^{1.1}$	$94.2^{0.3}$	$94.9^{2.3}$	$94.5^{0.4}$
5 20 1 3	99.5^{0}	95.2^{0}	95.1^{0}	95.2^{0}	94.8^{55}	$85.7^{9.6}$	95.9^{0}	98.9^{0}	$94.6^{3.4}$	$95.0^{3.8}$	$98.9^{0.2}$
5 20 4 1	88.5^{0}	95.8^{0}	93.0^{0}	93.8^{0}	95.7^{0}	95.1^{0}	49.5^{0}	$87.9^{1.4}$	$88.2^{0.6}$	$91.5^{3.1}$	$89.1^{2.3}$
5 20 4 2	95.5^{0}	95.7^{0}	$91.1^{0.5}$	$90.7^{0.5}$	$95.0^{9.6}$	$93.9^{9.7}$	96.3^{0}	91.8^{0}	$93.4^{0.7}$	$94.8^{2.1}$	95.1^{0}
5 20 4 3	98.6^{0}	95.1^{0}	78.8^{0}	74.7^{0}	$95.0^{0.2}$	$94.6^{0.4}$	$83.4^{8.5}$	$57.4^{3.3}$	$60.8^{7.3}$	$58.5^{3.3}$	$58.5^{8.1}$
20 5 1 1	89.8^{0}	96.9^{0}	92.0^{0}	92.5^{0}	95.2^{69}	94.4^{35}	95.2^{0}	$88.4^{1.1}$	$87.1^{1.4}$	$92.0^{2.8}$	$87.5^{4.2}$
20 5 1 2	95.0^{0}	95.9^{0}	92.3^{0}	92.8^{0}	95.4^{97}	84.3^{97}	95.2^{0}	$94.1^{1.6}$	$93.7^{0.3}$	$94.6^{1.9}$	$93.7^{3.6}$
20 5 1 3	99.0^{0}	95.5^{0}	92.5^{0}	93.1^{0}	92.5^{93}	74.7^{98}	95.8^{0}	$95.0^{0.9}$	$94.6^{4.3}$	$95.1^{2.1}$	$95.0^{7.2}$
20 5 4 1	88.9^{0}	96.9^{0}	84.4^{0}	88.8^{0}	95.4^{34}	94.8^{34}	61.6^{0}	$87.5^{3.3}$	$88.4^{0.9}$	$92.0^{1.6}$	$88.8^{4.2}$
20 5 4 2	95.3^{0}	96.1^{0}	91.0^{0}	90.4^{0}	93.4^{47}	94.4^{67}	96.0^{0}	$92.4^{6.8}$	$90.7^{0.6}$	$93.0^{1.7}$	$90.7^{5.8}$
20 5 4 3	97.9^{0}	95.2^{0}	91.4^{0}	82.2^{0}	89.0^{76}	93.8^{84}	$79.2^{8.3}$	$55.4^{4.5}$	56.8^{11}	$54.3^{4.9}$	54.7^{14}
5 100 1 1	91.6^{0}	94.9^{0}	95.0^{0}	95.0^{0}	94.8^{51}	94.8^{0}	94.8^{0}	$91.6^{2.0}$	$91.5^{0.2}$	$93.5^{2.3}$	$91.7^{2.0}$
5 100 1 2	94.8^{0}	94.9^{0}	95.0^{0}	95.0^{0}	94.8^{51}	94.5^{0}	95.0^{0}	$97.0^{4.3}$	$94.6^{0.2}$	$95.3^{1.7}$	$94.9^{6.2}$
5 100 1 3	99.4^{0}	94.8^{0}	95.1^{0}	95.1^{0}	94.5^{52}	94.0^{0}	94.9^{0}	94.2^{0}	$94.4^{0.1}$	$94.5^{0.8}$	$94.5^{1.5}$
5 100 4 1	88.3^{0}	94.9^{0}	83.2^{0}	84.1^{0}	95.0^{0}	94.8^{0}	1.5^{0}	$88.4^{2.8}$	87.9^{0}	$91.8^{4.8}$	$88.5^{0.8}$
5 100 4 2	95.1^{0}	95.3^{0}	64.9^{0}	64.2^{0}	95.1^{0}	94.7^{0}	95.5^{0}	$94.5^{2.3}$	$94.4^{0.1}$	$95.3^{1.9}$	$94.4^{0.6}$
5 100 4 3	98.2^{0}	95.0^{0}	0.11^{0}	0.10^{0}	95.0^{0}	94.8^{0}	3.9^{0}	$16.2^{0.1}$	$3.5^{3.5}$	$16.7^{0.6}$	$3.6^{5.2}$
100 5 1 1	90.5^{0}	96.5^{0}	92.4^{0}	92.5^{0}	NA^{100}	NA^{99}	94.9^{0}	$88.7^{0.2}$	$87.5^{5.4}$	$93.2^{0.2}$	$87.4^{6.2}$
100 5 1 2	94.8^{0}	95.8^{0}	92.0^{0}	92.2^{0}	NA^{100}	NA^{100}	94.8^{0}	$94.1^{3.4}$	$93.5^{0.3}$	$94.1^{1.8}$	$93.5^{6.4}$
100 5 1 3	98.9^{0}	95.4^{0}	91.9^{0}	92.4^{0}	NA^{100}	NA^{100}	94.9^{0}	$94.6^{0.6}$	94.2^{0}	$94.3^{1.6}$	$95.1^{2.0}$
100 5 4 1	80.7^{0}	97.1^{0}	46.5^{0}	67.6^{0}	91.1^{96}	92.7^{96}	4.4^{0}	$81.8^{3.6}$	$85.5^{3.9}$	$91.4^{0.4}$	$85.5^{6.0}$
100 5 4 2	95.4^{0}	96.2^{0}	75.6^{0}	67.6^{0}	NA^{100}	NA^{100}	95.1^{0}	$93.4^{5.8}$	$82.8^{0.7}$	$93.6^{1.6}$	$82.5^{9.5}$
100 5 4 3	94.0^{0}	95.6^{0}	89.4^{0}	54.9^{0}	NA^{100}	NA^{100}	0.20^{0}	1.1^{0}	$0.40^{0.4}$	$1.1^{2.0}$	$0.40^{5.0}$

Table 4: Relative Mean Squared Error (MSE) with common odds ratio assumption not holding and $\rho = 0.8$

K n_k Ψ S	L	$\tilde{L}$	L_{BT}	$\tilde{L}_{BT}$	GEE	ML	CML	MM1	MM3
5 20 1 1	$2.41^{0}_{0.0-}$	$2.56^{0}_{0.0-}$	$2.05^{0}_{0.0-}$	$2.15^{0}_{0.0-}$	2.71^{53}_{67+}	$2.58^{0.1}_{0.1-}$	$1.00^{0}_{0.0-}$	$2.48^{2.9}_{0.0-}$	$2.63^{4.1}_{0.2+}$
5 20 1 2	2.00^{0}_{18-}	2.00^{0}_{18-}	1.03^{0}_{19-}	1.00^{0}_{19-}	1.44^{73}_{63+}	$2.23^{3.7}_{18-}$	2.08^{0}_{17-}	$1.96^{0.7}_{19-}$	$2.46^{1.7}_{14-}$
5 20 1 3	1.23^{0}_{36+}	1.18^{0}_{36+}	1.05^{0}_{36+}	1.00^{0}_{36+}	1.57^{25}_{72+}	1.69^{15}_{34+}	5.11^{0}_{35+}	$3.51^{0.3}_{35+}$	$3.96^{0.8}_{35+}$
5 20 4 1	1.57^{0}_{10+}	1.70^{0}_{14+}	1.00^{0}_{18-}	1.00^{0}_{17-}	$1.76^{1.6}_{16+}$	$1.77^{1.6}_{16+}$	1.25^{0}_{35-}	$1.27^{0.5}_{12-}$	$1.88^{5.2}_{8.5+}$
5 20 4 2	2.44^{0}_{34+}	2.45^{0}_{34+}	1.00^{0}_{23-}	1.03^{0}_{27-}	$2.72^{0.1}_{39+}$	$2.70^{0.2}_{38+}$	3.29^{0}_{31+}	$2.72^{3.8}_{37+}$	$3.11^{1.9}_{35+}$
5 20 4 3	$1.01^{0}_{0.1-}$	$1.00^{0}_{2.0-}$	1.15^{0}_{33-}	1.26^{0}_{41-}	$1.03^{0}_{0.1+}$	$1.03^{0.1}_{0.0-}$	$23.3^{6.6}_{78+}$	$9.61^{6.5}_{63+}$	$11.6^{6.5}_{74+}$
20 5 1 1	$1.85^{0}_{0.0-}$	$2.25^{0}_{0.0-}$	$1.39^{0}_{0.0-}$	$1.53^{0}_{0.0-}$	2.46^{88}_{68+}	$2.35^{75}_{0.2+}$	$1.00^{0}_{0.0-}$	$1.48^{0.1}_{0.0-}$	$2.31^{2.4}_{0.0+}$
20 5 1 2	$1.28^{0}_{2.5+}$	$1.29^{0}_{2.6+}$	$1.05^{0}_{2.6+}$	$1.00^{0}_{2.6+}$	1.74^{84}_{68+}	$1.66^{86}_{1.9+}$	$1.32^{0}_{2.5+}$	$1.31^{0.8}_{2.4+}$	$1.47^{1.3}_{1.2+}$
20 5 1 3	$1.54^{0}_{2.2-}$	$1.26^{0}_{2.2-}$	$1.25^{0}_{2.2-}$	$1.00^{0}_{2.2-}$	1.64^{96}_{68+}	$1.16^{99}_{0.4-}$	$5.15^{0}_{2.2-}$	$2.47^{0.2}_{1.6-}$	$4.00^{0.4}_{1.8-}$
20 5 4 1	1.00^{0}_{14-}	$1.07^{0}_{1.3-}$	1.34^{0}_{62-}	1.29^{0}_{59-}	$1.14^{79}_{0.2+}$	$1.15^{79}_{0.3+}$	1.59^{0}_{68-}	1.03^{10}_{11-}	$1.06^{3.3}_{2.0-}$
20 5 4 2	$1.00^{0}_{0.6-}$	$1.01^{0}_{0.6-}$	1.14^{0}_{46-}	1.29^{0}_{56-}	$1.25^{89}_{8.7+}$	$1.19^{93}_{6.6+}$	$1.24^{0}_{0.1-}$	$1.11^{7.5}_{0.0-}$	$1.20^{1.5}_{1.1+}$
20 5 4 3	$1.11^{0}_{3.8+}$	$1.00^{0}_{3.8-}$	1.06^{0}_{38-}	1.65^{0}_{66-}	1.69^{90}_{21+}	$1.24^{94}_{0.2-}$	$17.1^{6.6}_{75+}$	$6.96^{5.7}_{61+}$	$8.96^{6.2}_{69+}$
5 100 1 1	2.03^{0}_{47+}	2.05^{0}_{47+}	1.42^{0}_{47+}	1.43^{0}_{47+}	2.41^{18}_{72+}	1.83^{0}_{41+}	1.00^{0}_{47+}	$1.39^{4.4}_{46+}$	$1.74^{3.8}_{34+}$
5 100 1 2	1.60^{0}_{17-}	1.60^{0}_{17-}	1.56^{0}_{17-}	1.56^{0}_{17-}	1.00^{69}_{63+}	1.63^{0}_{17-}	1.61^{0}_{17-}	$1.62^{0.7}_{17-}$	$1.58^{1.7}_{13-}$
5 100 1 3	$1.14^{0}_{9.3+}$	$1.13^{0}_{9.3+}$	$1.01^{0}_{9.3+}$	$1.00^{0}_{9.3+}$	1.40^{39}_{68+}	$1.21^{0}_{6.6+}$	$4.59^{0}_{9.2+}$	$1.03^{4.5}_{9.3+}$	$4.06^{0.8}_{16+}$
5 100 4 1	1.48^{0}_{26+}	1.54^{0}_{28+}	$1.00^{0}_{0.5-}$	$1.02^{0}_{0.2-}$	1.67^{0}_{32+}	1.51^{0}_{26+}	5.13^{0}_{88-}	$1.34^{3.8}_{1.2+}$	$1.50^{4.4}_{26+}$
5 100 4 2	$1.00^{0}_{1.7-}$	$1.00^{0}_{1.7-}$	2.95^{0}_{81-}	2.98^{0}_{81-}	$1.00^{0}_{0.6-}$	$1.01^{0}_{0.6-}$	$1.15^{0}_{1.1-}$	$1.13^{1.8}_{0.6-}$	$1.02^{1.6}_{0.1-}$
5 100 4 3	$1.00^{0}_{0.1-}$	$1.00^{0}_{0.6-}$	8.37^{0}_{94-}	8.60^{0}_{94-}	$1.02^{0}_{0.4+}$	$1.05^{0}_{0.5+}$	81.3^{0}_{92+}	$35.5^{0.1}_{91+}$	$36.8^{0.6}_{90+}$
100 5 1 1	$1.90^{0}_{1.5+}$	$2.31^{0}_{1.5+}$	$1.41^{0}_{1.5+}$	$1.55^{0}_{1.5+}$	NA^{100}	NA^{100}	$1.00^{0}_{1.5+}$	$1.87^{3.5}_{1.1+}$	$2.32^{0.7}_{1.2+}$
100 5 1 2	$1.38^{0}_{3.7-}$	$1.38^{0}_{3.7-}$	$1.06^{0}_{3.8-}$	$1.00^{0}_{3.7-}$	NA^{100}	NA^{100}	$1.39^{0}_{3.8-}$	$1.35^{3.5}_{5.1-}$	$1.42^{1.1}_{3.7-}$
100 5 1 3	$1.76^{0}_{0.0+}$	$1.44^{0}_{0.0+}$	$1.27^{0}_{0.0+}$	$1.00^{0}_{0.0+}$	NA^{100}	NA^{100}	$5.32^{0}_{0.0+}$	$4.01^{0.4}_{0.0+}$	$4.14^{0.8}_{0.0+}$
100 5 4 1	1.13^{0}_{26-}	$1.09^{0}_{3.1+}$	2.33^{0}_{72-}	1.63^{0}_{54-}	NA^{99}	NA^{99}	6.05^{0}_{92-}	$1.00^{5.5}_{5.3-}$	$1.10^{0.6}_{2.4+}$
100 5 4 2	$1.00^{0}_{0.8-}$	$1.01^{0}_{0.8-}$	1.91^{0}_{61-}	2.37^{0}_{70-}	NA^{99}	NA^{100}	$1.19^{0}_{0.4-}$	$1.00^{5.6}_{0.1-}$	$1.04^{1.5}_{0.3+}$
100 5 4 3	1.66^{0}_{34+}	$1.00^{0}_{1.8-}$	1.83^{0}_{61-}	4.48^{0}_{86-}	NA^{100}	NA^{100}	72.0^{0}_{92+}	$29.0^{0.2}_{91+}$	$30.2^{0.6}_{91+}$

Table 5: Coverage with common odds ratio assumption not holding with $\rho = 0.8$

K n_k Ψ S	L	$\tilde{L}$	L_{perc}	$\tilde{L}_{perc}$	GEE	ML	CML	MM1	MM2	MM3	MM4
5 20 1 1	88.6^{0}	95.6^{0}	94.8^{0}	94.8^{0}	95.5^{53}	$95.1^{0.1}$	95.3^{0}	$87.7^{2.9}$	$87.7^{0.5}$	$94.3^{4.1}$	90.5^{11}
5 20 1 2	90.5^{0}	90.9^{0}	90.3^{0}	90.4^{0}	96.0^{73}	$88.0^{3.7}$	91.2^{0}	$89.8^{0.7}$	$89.5^{0.3}$	$94.1^{1.7}$	$90.2^{5.4}$
5 20 1 3	97.3^{0}	88.0^{0}	88.1^{0}	88.1^{0}	84.6^{25}	75.0^{15}	89.5^{0}	$87.7^{0.3}$	$87.5^{4.5}$	$91.9^{0.8}$	88.8^{10}
5 20 4 1	91.1^{0}	95.2^{0}	91.9^{0}	92.1^{0}	$94.8^{1.6}$	$94.3^{1.6}$	85.5^{0}	$87.2^{0.5}$	$89.9^{0.9}$	$94.5^{5.2}$	91.2^{15}
5 20 4 2	91.0^{0}	91.5^{0}	90.5^{0}	89.4^{0}	$89.3^{0.1}$	$88.2^{0.2}$	94.7^{0}	$87.3^{3.8}$	$85.1^{0.3}$	$91.7^{1.9}$	$85.2^{6.0}$
5 20 4 3	98.8^{0}	94.3^{0}	87.8^{0}	85.3^{0}	95.2^{0}	$94.2^{0.1}$	$78.7^{6.6}$	$55.4^{2.9}$	$57.5^{4.8}$	$60.8^{4.7}$	$56.8^{8.9}$
20 5 1 1	90.9^{0}	96.8^{0}	92.2^{0}	92.7^{0}	95.4^{88}	94.2^{75}	95.4^{0}	$92.0^{0.1}$	$88.1^{0.4}$	$93.1^{2.4}$	89.3^{20}
20 5 1 2	95.1^{0}	95.9^{0}	92.0^{0}	92.5^{0}	94.3^{84}	82.6^{86}	95.2^{0}	$94.0^{0.8}$	$93.7^{0.2}$	$95.0^{1.3}$	93.7^{19}
20 5 1 3	99.2^{0}	95.7^{0}	92.9^{0}	93.5^{0}	95.8^{96}	77.0^{99}	96.1^{0}	$96.4^{0.2}$	$94.9^{3.6}$	$95.9^{0.4}$	95.3^{13}
20 5 4 1	87.5^{0}	96.1^{0}	69.7^{0}	72.5^{0}	94.5^{79}	94.0^{79}	64.4^{0}	85.9^{10}	$89.5^{0.4}$	$92.2^{3.3}$	89.7^{23}
20 5 4 2	94.7^{0}	95.5^{0}	81.2^{0}	76.1^{0}	94.9^{89}	94.7^{93}	95.3^{0}	$92.0^{7.5}$	$92.7^{0.3}$	$94.2^{1.5}$	92.6^{18}
20 5 4 3	98.5^{0}	93.8^{0}	84.7^{0}	69.4^{0}	93.2^{90}	91.7^{94}	$87.8^{6.6}$	$64.7^{3.8}$	65.6^{10}	$66.5^{5.0}$	63.9^{18}
5 100 1 1	77.0^{0}	86.1^{0}	86.0^{0}	86.0^{0}	83.1^{18}	87.9^{0}	86.1^{0}	$82.2^{4.4}$	76.8^{0}	$89.5^{3.8}$	$81.7^{5.6}$
5 100 1 2	92.5^{0}	92.5^{0}	92.7^{0}	92.8^{0}	97.4^{69}	92.1^{0}	92.6^{0}	$92.3^{0.7}$	$92.2^{0.3}$	$95.7^{1.7}$	$92.9^{1.9}$
5 100 1 3	99.4^{0}	94.1^{0}	94.4^{0}	94.4^{0}	92.1^{39}	93.2^{0}	94.4^{0}	$99.4^{4.5}$	93.4^{0}	$98.1^{0.8}$	$95.0^{0.7}$
5 100 4 1	83.7^{0}	90.5^{0}	92.9^{0}	92.8^{0}	89.8^{0}	91.0^{0}	20.2^{0}	$83.5^{3.8}$	$80.8^{0.1}$	$93.3^{4.4}$	$86.6^{6.7}$
5 100 4 2	91.2^{0}	91.1^{0}	41.8^{0}	41.2^{0}	91.3^{0}	91.0^{0}	91.5^{0}	$88.6^{1.8}$	$90.8^{0.1}$	$95.4^{1.6}$	$91.8^{1.5}$
5 100 4 3	98.5^{0}	94.6^{0}	4.4^{0}	4.0^{0}	95.1^{0}	94.4^{0}	1.2^{0}	$3.4^{0.1}$	$0.98^{1.3}$	$16.6^{0.6}$	$3.6^{2.3}$
100 5 1 1	90.8^{0}	96.7^{0}	92.3^{0}	92.6^{0}	NA^{100}	NA^{100}	95.4^{0}	$89.3^{3.5}$	$87.7^{4.7}$	$93.7^{0.7}$	88.1^{21}
100 5 1 2	94.6^{0}	95.6^{0}	91.5^{0}	91.9^{0}	NA^{100}	NA^{100}	94.6^{0}	$93.6^{3.5}$	$92.8^{0.6}$	$94.2^{1.1}$	92.8^{21}
100 5 1 3	98.9^{0}	95.6^{0}	92.7^{0}	93.1^{0}	NA^{100}	NA^{100}	95.4^{0}	$94.8^{0.4}$	94.9^{0}	$94.8^{0.8}$	$94.9^{4.5}$
100 5 4 1	84.5^{0}	96.4^{0}	55.8^{0}	74.0^{0}	NA^{99}	NA^{99}	7.7^{0}	$86.5^{5.5}$	$82.2^{4.0}$	$92.3^{0.6}$	82.5^{26}
100 5 4 2	94.9^{0}	95.9^{0}	71.4^{0}	62.0^{0}	NA^{99}	NA^{100}	94.5^{0}	$93.6^{5.6}$	$84.0^{0.9}$	$94.4^{1.5}$	84.0^{24}
100 5 4 3	95.1^{0}	94.1^{0}	68.7^{0}	25.7^{0}	NA^{100}	NA^{100}	0.84^{0}	$3.6^{0.1}$	$1.3^{0.4}$	$3.3^{0.6}$	$1.4^{5.8}$

Table 6: Power for various configurations and $\Psi = 2$, H indicates whether the common odds ratio assumption holds. H =T when it holds and H =F when it does not hold (with $\rho = 0.8$).

K n_k S H	L	$\tilde{L}$	L_{perc}	$\tilde{L}_{perc}$	GEE	ML	CML	MM1	MM2	MM3	MM4
5 20 1 T	62.4^{0}	42.1^{0}	45.7^{0}	45.7^{0}	$45.5^{3.8}$	45.0^{0}	44.0^{0}	$63.1^{1.4}$	$63.0^{0.4}$	$50.0^{4.5}$	57.8^{11}
5 20 2 T	63.7^{0}	62.5^{0}	64.3^{0}	64.3^{0}	$64.2^{1.3}$	65.6^{0}	61.9^{0}	$62.4^{2.0}$	$65.6^{0.3}$	$56.9^{2.5}$	$64.3^{4.5}$
5 20 3 T	71.1^{0}	90.4^{0}	90.8^{0}	90.6^{0}	$90.9^{0.1}$	$92.9^{1.4}$	$89.7^{0.3}$	$74.1^{0.1}$	$91.3^{1.6}$	$88.6^{2.8}$	$90.8^{6.8}$
5 20 1 F	73.0^{0}	59.0^{0}	61.8^{0}	61.7^{0}	$60.8^{1.8}$	$59.9^{0.2}$	60.4^{0}	$69.3^{3.1}$	$73.7^{0.8}$	$57.5^{4.7}$	67.8^{15}
5 20 2 F	36.2^{0}	35.0^{0}	37.0^{0}	36.8^{0}	$38.3^{7.9}$	$39.2^{2.3}$	34.9^{0}	$37.7^{2.0}$	$38.3^{0.2}$	$28.7^{2.1}$	$37.2^{6.7}$
5 20 3 F	56.8^{0}	79.9^{0}	80.0^{0}	79.7^{0}	$79.7^{0.5}$	$85.4^{5.6}$	$77.8^{0.4}$	$56.1^{0.2}$	80.5^{17}	$69.1^{1.2}$	76.2^{21}
20 5 1 T	56.0^{0}	37.5^{0}	52.7^{0}	51.7^{0}	46.1^{37}	46.1^{35}	43.9^{0}	$58.9^{3.1}$	$60.9^{0.9}$	$48.5^{2.0}$	57.9^{24}
20 5 2 T	59.8^{0}	55.6^{0}	67.1^{0}	66.2^{0}	61.6^{75}	73.7^{89}	58.2^{0}	$62.6^{7.7}$	$63.9^{0.2}$	$58.2^{1.6}$	63.8^{19}
20 5 3 T	68.1^{0}	85.2^{0}	89.7^{0}	89.1^{0}	88.0^{82}	77.7^{94}	$85.0^{0.5}$	$77.4^{0.1}$	$86.1^{4.6}$	$85.2^{0.7}$	84.3^{13}
20 5 1 F	53.4^{0}	34.6^{0}	50.0^{0}	48.9^{0}	44.9^{90}	45.5^{90}	41.4^{0}	$56.3^{4.0}$	$58.2^{0.6}$	$45.0^{2.5}$	54.8^{20}
20 5 2 F	58.2^{0}	54.2^{0}	65.2^{0}	64.2^{0}	58.9^{94}	65.0^{97}	56.8^{0}	$60.6^{5.8}$	$62.4^{0.2}$	$57.2^{1.6}$	62.1^{16}
20 5 3 F	68.4^{0}	84.6^{0}	89.4^{0}	88.8^{0}	89.0^{88}	78.7^{96}	$84.0^{0.6}$	$86.9^{0.8}$	$85.4^{4.2}$	$82.7^{2.9}$	82.8^{12}

Table 7: Merck Data: The estimates from the standard MH L (1st row), the new MH $\tilde{L}$ (2nd row), GEE (3rd row), CML (4th row), MM1 (5th row), and MM3 (6th row) methods along with their standard errors in parentheses

	Treatment				
Methods	1	2	3	4	Combined
standard MH L	0.850 (0.444)	0.890 (0.412)	0.296 (0.385)	0.072 (0.379)	0.419 (0.184)
New MH $\tilde{L}$	0.621 (0.323)	0.856 (0.316)	0.260 (0.180)	0.000 (0.336)	0.400 (0.114)
GEE	NA	NA	NA	NA	0.428 (0.123)
CML	1.705 (0.768)	0.872 (0.760)	1.099 (0.816)	0.118 (0.486)	1.025 (0.311)
MM1	2.698 (0.897)	1.479 (0.514)	0.716 (0.610)	0.107 (0.416)	0.857 (0.251)
MM3	NA	1.500 (0.534)	0.716 (0.610)	0.108 (0.417)	0.859 (0.260)

*

A Appendix

A.1 Non-dual Consistency of $\hat{\Psi}_{xyy}$

Under the *Sparse-Data Limiting Model* (Limiting Model II) we can write

$$\begin{aligned}\hat{\Psi}_{xy} - \Psi_{xy} &= \frac{C_{xy} - \Psi_{xy} C_{yx}}{C_{yx}} = \frac{\sum_{k=1}^{K} c_{xy|k} - \Psi_{xy} c_{yx|k}}{\sum_{k=1}^{K} c_{yx|k}} \\ &= \frac{(\sum_{k=1}^{K} c_{xy|k} - \Psi_{xy} c_{yx|k})/K}{\sum_{k=1}^{K} c_{yx|k}/K} = \frac{(C_{xy} - \Psi_{xy} C_{yx})/K}{C_{yx}/K} \\ &= \frac{\sum_{k=1}^{K} \omega_{xy|k}/K}{\sum_{k=1}^{K} c_{yx|k}/K} = \frac{\Omega_{xy}/K}{C_{yx}/K}\end{aligned}$$

with $\omega_{xy|k} = c_{xy|k} - \Psi_{xy} c_{yx|k}$ and $\Omega = \sum_k \omega_k$.

The term $c_{xy|k}$ is a bounded random variable under limiting model II, hence the variance of C_{xy} is $o(K^2)$ and Chebyshev's weak law of large numbers (CWLLN) implies $\frac{1}{K}(\Omega_{xy} - \mathbb{E}\Omega_{xy}) \to_p 0$.

We assume $\mathbf{n}_k = (n^{00}_{xy|k}, n^{01}_{xy|k}, n^{10}_{xy|k}, n^{11}_{xy|k})$ follows a multinomial distribution with pa-

rameters n_k and $\boldsymbol{\pi}_k = (\pi^{00}_{xy|k}, \pi^{01}_{xy|k}, \pi^{10}_{xy|k}, \pi^{11}_{xy|k})$. It follows:

$$\mathbb{E}n_{x|k}n_{y|k} = \mathbb{E}(n^{10}_{xy|k} + n^{11}_{xy|k})(n^{01}_{xy|k} + n^{11}_{xy|k}) = n_k[n'_k\pi_{x|k}\pi_{y|k} + \pi^{11}_{xy|k}]$$

with $n'_k = n_k - 1$.

Therefore,

$$\begin{aligned}
\mathbb{E}c_{xy|k} &= \frac{1}{n_k}\mathbb{E}n_{x|k}\bar{n}_{y|k} = \frac{1}{n_k}\mathbb{E}n_{x|k}(n_k - n_{y|k}) = \frac{1}{n_k}\left(n_k\mathbb{E}n_{x|k} - \mathbb{E}n_{x|k}n_{y|k}\right) \\
&= \frac{1}{n_k}\left(n_k^2\pi_{x|k} - n_k\left[n'_k\pi_{x|k}\pi_{y|k} + \pi^{11}_{xy|k}\right]\right) \\
&= n_k\pi_{x|k} - n_k\pi_{x|k}\pi_{y|k} + \pi_{x|k}\pi_{y|k} - \pi^{11}_{xy|k} \\
&= n_k\pi_{x|k}\bar{\pi}_{y|k} + (\pi_{x|k}\pi_{y|k} - \pi^{11}_{xy|k})
\end{aligned}$$

We conclude

$$\begin{aligned}
&\mathbb{E}\Omega_{xy} = \mathbb{E}(C_{xy} - \Psi C_{yx}) \\
&= \sum_k \left\{ n_k\pi_{x|k}\bar{\pi}_{y|k} + (\pi_{x|k}\pi_{y|k} - \pi^{11}_{xy|k}) - \Psi_{xy}\left(n_k\pi_{y|k}\bar{\pi}_{x|k} + (\pi_{y|k}\pi_{x|k} - \pi^{11}_{xy|k})\right)\right\} \\
&= \sum_k \left\{ n_k\pi_{x|k}\bar{\pi}_{y|k} - n_k\pi_{x|k}\bar{\pi}_{y|k} + (\pi_{x|k}\pi_{y|k} - \pi^{11}_{xy|k})(1 - \Psi_{xy})\right\} \\
&= (1 - \Psi_{xy})\sum_k(\pi_{x|k}\pi_{y|k} - \pi^{11}_{xy|k}).
\end{aligned}$$

Therefore $\frac{1}{K}\Omega_{xy}$ converges to a non-zero constant, i.e. $\lim_{K\to\infty}\frac{1}{K}\mathbb{E}\Omega_{xy}$, unless items are independent or $\Psi_{xy} = 1$. Since $\hat{\Psi}_{xy} - \Psi_{xy}$ does not converge to 0 in probability, $\hat{\Psi}_{xy}$ is not consistent for the general case of b).

Under *Large-Stratum Limiting Model* (Limiting Model I), let $N = \sum_k n_k$ and $\frac{n_k}{N} \to \alpha_k > 0$ as $N \to \infty$. Then,

$$\begin{aligned}
\hat{\Psi}_{xy} =& \frac{\sum_{k=1}^K n_{x|k}\bar{n}_{y|k}/n_k}{\sum_{k=1}^K n_{y|k}\bar{n}_{x|k}/n_k} = \frac{\sum_k \frac{1}{n_kN}n_{x|k}\bar{n}_{y|k}}{\sum_k \frac{1}{n_kN}n_{y|k}\bar{n}_{x|k}} = \frac{\sum_k \frac{n_k}{N}\frac{n_{x|k}}{n_k}\frac{\bar{n}_{y|k}}{n_k}}{\sum_k \frac{n_k}{N}\frac{n_{y|k}}{n_k}\frac{\bar{n}_{x|k}}{n_k}} \\
\xrightarrow[p]{N\to\infty}& \frac{\sum_k \alpha_k\pi_{x|k}\bar{\pi}_{y|k}}{\sum_k \alpha_k\pi_{y|k}\bar{\pi}_{x|k}} = \Psi_{xy}\frac{\sum_k \alpha_k\pi_{y|k}\bar{\pi}_{x|k}}{\sum_k \alpha_k\pi_{y|k}\bar{\pi}_{x|k}} = \Psi_{xy},
\end{aligned}$$

by $\pi_x\bar{\pi}_y = \Psi_{xy}\bar{\pi}_x\pi_y$, that is, the consistency holds under limiting model I. The ordinary MH estimator is not only dually consistent under independence of items, but also when $\Psi_{xy} = 1$ and even when items are dependent.

A.2 Dual Consistency of $\tilde{\Psi}_{xy}$

Sparse-Data Limiting Model (Limiting Model II):

As before, we can write

$$\begin{aligned}\tilde{\Psi}_{xy} - \Psi_{xy} &= \frac{(\sum_{k=1}^{K} \tilde{c}_{xy|k} - \Psi_{xy}\tilde{c}_{yx|k})/K}{\sum_{k=1}^{K} \tilde{c}_{yx|k}/K} = \frac{(\tilde{C}_{xy} - \Psi_{xy}\tilde{C}_{yx})/K}{\tilde{C}_{yx}/K} \\ &= \frac{(\sum_{k=1}^{K} \tilde{\omega}_{xy|k})/K}{\sum_{k=1}^{K} \tilde{c}_{yx|k}/K} = \frac{\tilde{\Omega}_{xy}/K}{\tilde{C}_{yx}/K}\end{aligned} \tag{10}$$

with $\tilde{\omega}_{xy|k} = \tilde{c}_{xy|k} - \Psi_{xy}\tilde{c}_{yx|k}$ and $\tilde{\Omega} = \sum_k \tilde{\omega}_k$. We have

$$\begin{aligned}\mathbb{E}\tilde{c}_{xy|k} &= \mathbb{E}(n_{x|k}\bar{n}_{y|k} - n^{10}_{xy|k})/n_k = \frac{1}{n_k}\left(\mathbb{E}n_{x|k}(n_k - n_{y|k}) - \mathbb{E}n^{10}_{xy|k}\right) \\ &= \frac{1}{n_k}\left(n_k\mathbb{E}n_{x|k} - \mathbb{E}n_{x|k}n_{y|k} - \mathbb{E}n^{10}_{xy|k}\right) \\ &= \frac{1}{n_k}\left(n_k^2\pi_{x|k} - n_k(n'_k\pi_{x|k}\pi_{y|k} + \pi^{11}_{xy|k}) - n_k\pi^{10}_{xy|k}\right) \\ &= \frac{1}{n_k}\left(n_k n'_k(\pi_{x|k} - \pi_{x|k}\pi_{y|k}) + n_k(\pi_{x|k} - \pi^{11}_{xy|k} - \pi^{10}_{xy|k})\right) \\ &= n'_k\pi_{x|k}\bar{\pi}_{y|k} + \frac{n_k}{n_k}(\pi_{x|k} - \pi_{x|k}) = n'_k\pi_{x|k}\bar{\pi}_{y|k},\end{aligned}$$

hence, $\mathbb{E}\tilde{\Omega}_{xy} = \mathbb{E}(\tilde{C}_{xy} - \Psi_{xy}\tilde{C}_{yx}) = 0$. This results holds also for the special case of independence between items. We apply Chebyshev's weak law of large numbers and find

$$\tilde{C}_{xy}/K = \sum_{k=1}^{K} \tilde{c}_{xy|k}/K \overset{K\to\infty}{\longrightarrow}_p \lim_{K\to\infty} \sum_{k=1}^{K} \mathbb{E}(\tilde{c}_{xy|k})/K = \lim_{K\to\infty} \mathbb{E}\tilde{C}_{xy}/K. \tag{11}$$

It follows from equation (10) and by applying Chebyshev's weak law of large numbers to the numerator together with $\mathbb{E}\tilde{\Omega}_{xy} = 0$ and equation (11), that the new estimator $\tilde{\Psi}_{xy}$ is consistent under limiting model II, in contrast to $\hat{\Psi}_{xy}$.

Large-Stratum Limiting Model (Limiting Model I):

$$\begin{aligned}\tilde{C}_{xy|k}/N &= \sum_{k=1}^{K} \tilde{c}_{xy|k}/N = \sum_{k=1}^{K} (n_{x|k}\bar{n}_{y|k} - n^{10}_{xy|k})/(n_k N) \\ &= \sum_{k=1}^{K} \frac{n_k^2}{n_k N}\frac{n_{x|k}}{n_k}\frac{\bar{n}_{y|k}}{n_k} - \frac{n_k}{n_k N}\frac{n^{10}_{xy|k}}{n_k} \\ &\overset{N\to\infty}{\longrightarrow}_p \sum_{k=1}^{K} \alpha_k\pi_{x|k}\bar{\pi}_{y|k} - 0\cdot\pi^{10}_{xy|k} = \sum_{k=1}^{K} \alpha_k\pi_{x|k}\bar{\pi}_{y|k}.\end{aligned}$$

Now

$$\lim_{N\to\infty}\tilde{\Psi}_{xy}=\frac{\lim_N \mathbb{E}\tilde{C}_{xy}/N}{\lim_N \mathbb{E}\tilde{C}_{yx}/N}=\Psi_{xy}\frac{\lim_N \mathbb{E}\tilde{C}_{yx}/N}{\lim_N \mathbb{E}\tilde{C}_{yx}/N}=\Psi_{xy}.$$

Thus, $\tilde{\Psi}_{xy}$ is indeed dually consistent.

A.3 Dual Consistency of $\tilde{U}_{xyy}$

For the sparse-data limiting model, we obtain the following asymptotic variance:

$$\begin{aligned}\lim_{K\to\infty} K\cdot\mathrm{Var}^a(\tilde{L}_{xy}) &= \frac{1}{\Psi_{xy}^2}\lim_{K\to\infty} K\cdot\mathrm{Var}^a(\tilde{\Psi}_{xy})\\ &= \frac{1}{\Psi_{xy}^2}\frac{\lim_{K\to\infty} K\cdot\mathrm{Var}^a(\tilde{\Omega}_{xy}/K)}{[\lim_{M\to\infty}\sum_{k=1}^K \mathbb{E}\tilde{c}_{yx|k}/K]^2}\\ &= \frac{\lim_{K\to\infty}\sum_k \mathrm{Var}(\tilde{\omega}_{xy|k})/K}{[\lim_{K\to\infty}\sum_{k=1}^K \mathbb{E}\tilde{c}_{xy|k}/K]^2},\end{aligned} \tag{12}$$

where

$$\begin{aligned}&\mathrm{Var}(\tilde{\omega}_{xy|k})=\mathrm{Var}(\tilde{c}_{xy}-\Psi\tilde{c}_{yx})\\ &=\mathbb{E}(\tilde{c}_{xy}-\Psi\tilde{c}_{yx})^2-[\mathbb{E}(\tilde{c}_{xy}-\Psi\tilde{c}_{yx})]^2=\mathbb{E}(\tilde{c}_{xy}-\Psi\tilde{c}_{yx})^2\\ &=\mathbb{E}\tilde{c}_{xy}^2+\Psi^2\mathbb{E}\tilde{c}_{yx}^2-2\Psi\mathbb{E}\tilde{c}_{xy}\tilde{c}_{yx}\\ &=\frac{1}{n_k^2}\{\mathbb{E}(X_{x|k}\bar{X}_{y|k}-X_{xy|k}^{10})^2+\Psi^2\mathbb{E}(X_{y|k}\bar{X}_{x|k}-X_{yx|k}^{10})^2\\ &-2\Psi\mathbb{E}(X_{x|k}\bar{X}_{y|k}-X_{xy|k}^{10})(X_{y|k}\bar{X}_{x|k}-X_{yx|k}^{10})\}\end{aligned}$$

For convenience, we suppress subscript k and write $X_{st}:=X_{xy|k}^{st}$ for $s,\,t\in\{0,\,1\}$.

$$\begin{aligned}&=\frac{1}{n_k^2}\{(\mathbb{E}X_x^2\bar{X}_y^2+\mathbb{E}X_{10}^2-2\mathbb{E}X_x\bar{X}_yX_{10})+\Psi^2(\mathbb{E}X_y^2\bar{X}_x^2+\mathbb{E}X_{01}^2-2\mathbb{E}X_y\bar{X}_xX_{01})\\ &-2\Psi(\mathbb{E}X_xX_y\bar{X}_x\bar{X}_y-\mathbb{E}X_x\bar{X}_yX_{01}-\mathbb{E}X_y\bar{X}_xX_{10}+\mathbb{E}X_{10}X_{01})\}\\ &=\frac{1}{n_k^2}\{\mathbb{E}X_x^2\bar{X}_y^2+\mathbb{E}X_{10}^2-2\mathbb{E}X_x\bar{X}_yX_{10}+\Psi^2\mathbb{E}X_y^2\bar{X}_x^2+\Psi^2\mathbb{E}X_{01}^2-2\Psi^2\mathbb{E}X_y\bar{X}_xX_{01}\\ &-2\Psi\mathbb{E}X_xX_y\bar{X}_x\bar{X}_y+2\Psi\mathbb{E}X_x\bar{X}_yX_{01}+2\Psi\mathbb{E}X_y\bar{X}_xX_{10}-2\Psi\mathbb{E}X_{10}X_{01}\}.\end{aligned} \tag{13}$$

We define $N_3:=nn'n''n'''$, $N_2:=nn'n''$, $N_1:=nn'$, $N_0:=n$ with $n'=n-1$, $n''=n-2$ and $n'''=n-3$. Using the moment generating function of the multinomial distribution $(n,\,(p_1,\,p_2,\,p_3,\,\ldots))$, we can derive the following higher order moments (indices i, j and k

refer to different outcomes of the multinomial distribution):

$$\begin{aligned}
\mathbb{E}X_i &= N_0 p_i \\
\mathbb{E}X_i^2 &= N_1 p_i^2 + n p_i \\
\mathbb{E}X_i X_j &= N_1 p_i p_j \\
\mathbb{E}X_i^3 &= N_2 p_i^3 + 3N_1 p_i^2 + N_0 p_i \\
\mathbb{E}X_i^2 X_j &= N_2 p_i^2 p_j + N_1 p_i p_j \\
\mathbb{E}X_i X_j X_k &= N_2 p_i p_j p_k \\
\mathbb{E}X_i^4 &= N_3 p_i^4 + 6N_2 p_i^3 + 7N_1 p_i^2 + N_0 p_i \\
\mathbb{E}X_i^3 X_j &= N_3 p_i^3 p_j + 3N_2 p_i^2 p_j + N_1 p_i p_j \\
\mathbb{E}X_i^2 X_j^2 &= N_3 p_i^2 p_j^2 + N_2 (p_i^2 p_j + p_i p_j^2) + N_1 p_i p_j \\
\mathbb{E}X_i^2 X_j X_k &= N_3 p_i^2 p_j p_k + N_2 p_i p_j p_k \\
\mathbb{E}X_i X_j X_k X_l &= N_3 p_i p_j p_k p_l. \qquad (14)
\end{aligned}$$

For convenience, define $X_A := X_{10}$, $X_B := X_{01}$, $X_C := X_{11}$, $X_D := X_{00}$ to avoid confusion with the indices $s, t \in \{0, 1\}$, similarly for the π_{st}'s. Now we write n^2 and n as

$$\begin{aligned}
n^2 &= n''n''' + 5n'' + 4 = n'n'' + 3n' + 1 = nn' + n \\
n &= n''' + 3 = n'' + 2 = n' + 1,
\end{aligned}$$

hence,

$$\begin{aligned}
n^2 N_1 = n^2 n' &= N_3 + 5N_2 + 4N_1 & \qquad nN_2 = n^2 n' n'' &= N_3 + 3N_2 \\
n^2 N_0 = n^3 &= N_2 + 3N_1 + N_0 & \qquad nN_1 = n^2 n' &= N_2 + 2N_1 \\
n^2 &= N_1 + N_0 & \qquad nN_0 = n^2 &= N_1 + N_0. \qquad (15)
\end{aligned}$$

with $N_0 = n$, $N_1 = nn'$, $N_2 = nn'n''$ and $N_3 = nn'n''n'''$. Let$(\cdot)|_{N_i}$ denote the terms of $(\cdot)$ with factor N_i, for example $\mathbb{E}X_i^3 X_j|_{N_3} = p_i^3 p_j$. By applying (15) with (14), we can derive

the following higher moments as shown by Suesse [26]:

$$
\begin{aligned}
\mathbb{E}X_A^2 &= N_1\pi_A^2 + N_0\pi_A \\
\mathbb{E}X_B^2 &= N_1\pi_B^2 + N_0\pi_B \\
\mathbb{E}X_AX_B &= N_1\pi_A\pi_B \\
\mathbb{E}X_x\bar{X}_yX_A &= N_2\pi_x\bar{\pi}_y\pi_A + N_1\{2\pi_A^2 + \pi_A - \pi_A\pi_B\} + N_0\pi_A \\
\mathbb{E}X_x\bar{X}_yX_B &= N_2\pi_x\bar{\pi}_y\pi_B + N_1\pi_A\pi_B \\
\mathbb{E}\bar{X}_xX_yX_B &= N_2\pi_y\bar{\pi}_x\pi_B + N_1\{2\pi_B^2 + \pi_B - \pi_B\pi_A\} + N_0\pi_B \\
\mathbb{E}\bar{X}_xX_yX_A &= N_2\pi_y\bar{\pi}_x\pi_A + N_1\pi_B\pi_A \\
\mathbb{E}X_x^2\bar{X}_y^2|_{N_3} &= \pi_x^2\bar{\pi}_y^2 \\
\mathbb{E}X_x^2\bar{X}_y^2|_{N_2} &= \pi_x\bar{\pi}_y(1 - \pi_B + 5\pi_A) \\
\mathbb{E}X_x^2\bar{X}_y^2|_{N_1} &= \pi_x\bar{\pi}_y + 4\pi_A^2 + 2\pi_A - 2\pi_A\pi_B \\
\mathbb{E}X_x^2\bar{X}_y^2|_{N_0} &= \pi_A \\
\mathbb{E}X_y^2\bar{X}_x^2|_{N_3} &= \pi_y^2\bar{\pi}_x^2 \\
\mathbb{E}X_y^2\bar{X}_x^2|_{N_2} &= \pi_y\bar{\pi}_x(1 - \pi_A + 5\pi_B) \\
\mathbb{E}X_y^2\bar{X}_x^2|_{N_1} &= \pi_y\bar{\pi}_x + 4\pi_B^2 + 2\pi_B - 2\pi_A\pi_B \\
\mathbb{E}X_y^2\bar{X}_x^2|_{N_0} &= \pi_B \\
\mathbb{E}X_xX_y\bar{X}_x\bar{X}_y|_{N_3} &= \pi_x\pi_y\bar{\pi}_x\bar{\pi}_y \\
2 \times \mathbb{E}X_xX_y\bar{X}_x\bar{X}_y|_{N_2} &= (\pi_x\bar{\pi}_y + \pi_y\bar{\pi}_x)(2\pi_A + 2\pi_B + 1) - 2(\pi_A - \pi_B)^2 - (\pi_A + \pi_B) \\
\mathbb{E}X_xX_y\bar{X}_x\bar{X}_y|_{N_1} &= \pi_x\bar{\pi}_y - \pi_A = \pi_y\bar{\pi}_x - \pi_B \\
\mathbb{E}X_xX_y\bar{X}_x\bar{X}_y|_{N_0} &= 0.
\end{aligned} \tag{16}
$$

Finally we are able to compute (13) by using (16).

$$
\begin{aligned}
\mathrm{Var}(\tilde{\omega}_{xy|k}) = &\frac{N_2}{n^2}\Psi\{(\pi_A + \pi_B) - (\pi_A - \pi_B)^2\} \\
&+ \frac{N_1}{n^2}\{\pi_A^2 + \Psi^2\pi_B^2 + \Psi(\pi_A + \pi_B + 2\pi_A\pi_B)\}.
\end{aligned}
$$

The sparse data limiting variance is obtained by inserting $\mathrm{Var}(\tilde{\omega}_{xy|k})$ into equation (12).

For the large-stratum limiting model, we obtain the following asymptotic variance:

$$
\lim_{N\to\infty} N \cdot \mathrm{Var}^a(\tilde{L}_{xy}) = \frac{\lim_{N\to\infty}\sum_k \mathrm{Var}^a(\tilde{\omega}_{xy|k})/N}{[\lim_{N\to\infty}\sum_{k=1}^K \mathbb{E}\tilde{c}_{xy|k}/N]^2}.
$$

By the delta method, the large-stratum limiting variance is

$$
\lim_{N\to\infty}\frac{1}{N}\sum_k \mathrm{Var}^a(\tilde{\omega}_{xy}) = \sum_k \alpha_k\{\pi_x\bar{\pi}_x + \pi_y\bar{\pi}_y + 2(\pi_x\pi_y - \pi_C)\}.
$$

Now we can show that the estimator $\tilde{U}_{xyy}$ converges under both limiting situations to the corresponding asymptotic variance.

B Appendix

Let $\boldsymbol{\beta}_k := (\beta_{xk}, \beta_{yk})^T$ be the vector of random effects and F_k the distribution of $\boldsymbol{\beta}_k$ with covariance $\boldsymbol{\Sigma} = \begin{pmatrix} \sigma_x^2 & \rho\sigma_x\sigma_y \\ \rho\sigma_x\sigma_y & \sigma_y^2 \end{pmatrix}$ and mean $\mathbf{0}$.

First, we re-express the model

$$\begin{aligned} \text{logit}(\pi_{x|k}) &= \alpha_k + \beta_{xk} + \beta_x \\ \text{logit}(\pi_{y|k}) &= \alpha_k + \beta_{yk} + \beta_y \end{aligned}$$

for any two items as

$$\begin{aligned} \text{logit}(\pi_{x|k}) &= \tilde{\alpha}_k + \tilde{\beta}_{xk} + \beta_x \\ \text{logit}(\pi_{y|k}) &= \tilde{\alpha}_k + \tilde{\beta}_{yk} + \beta_y \end{aligned}$$

with $\tilde{\alpha}_k = \alpha_k - \lambda_1\beta_{xk} - \lambda_2\beta_{yk}$, $\tilde{\beta}_{xk} = (\lambda_1 + 1)\beta_{xk} + \lambda_2\beta_{yk}$ and $\tilde{\beta}_{yk} = \lambda_1\beta_{xk} + (\lambda_2 + 1)\beta_{yk}$. The two models are identical and imply the same probabilities $\pi_{x|k}$ and $\pi_{y|k}$, the same $\beta_{xyk} = \beta_{xk} - \beta_{yk} = \tilde{\beta}_{xk} - \tilde{\beta}_{yk}$ and $\beta_{xy} = \beta_x - \beta_y$ due to unchanged β_x and β_y. The random effect distribution of $\tilde{\boldsymbol{\beta}}_k$ also has mean $\mathbf{0}$ but covariance $\tilde{\boldsymbol{\Sigma}} \neq \boldsymbol{\Sigma}$. The elements of $\tilde{\boldsymbol{\Sigma}}$ are denoted by $\tilde{\sigma}_x^2$, $\tilde{\sigma}_y^2$ and $\tilde{\rho}_{xy}\tilde{\sigma}_x\tilde{\sigma}_y$. We choose λ_1 and λ_2 to have $\tilde{\sigma}_x = \tilde{\sigma}_y$ and $\tilde{\rho}_{xy} = 0$. The solutions of this problem for λ_1 are

$$\lambda_1 = \frac{-\sigma_x^2\sigma_y^2(1 + \rho_{xy}) \pm (\sigma_x^2 - \rho_{xy}\sigma_x\sigma_y)\sqrt{\sigma_x^2\sigma_y^2(1 + \rho_{xy})}}{2}$$

and the corresponding solutions for λ_2 are

$$\lambda_2 = \frac{-\sigma_x^2\sigma_y^2(1 + \rho_{xy}) \pm (\sigma_y^2 - \rho_{xy}\sigma_x\sigma_y)\sqrt{\sigma_x^2\sigma_y^2(1 + \rho_{xy})}}{2},$$

which imply a variance of $\tilde{\sigma}_x^2 = \tilde{\sigma}_y^2 = \{\sigma_x^2 + \sigma_y^2 - 2\rho_{xy}\sigma_x\sigma_y\}/2 = \text{Var}(\beta_{xk} - \beta_{yk})/2$.

Without the loss of generality, assume $\sigma_x^2 = \sigma_y^2$ and $\rho_{xy} = 0$.

In Section A of the Appendix, we have shown the in- and consistency of the MH estimators L and $\tilde{L}$ under limiting model II. In this situation, the ordinary MH estimator $\hat{\Psi}$ converges to

$$\Psi^\infty := \frac{\lim_{K\to\infty} \sum_k \mathbb{E}c_{xy|k}/K}{\lim_{K\to\infty} \sum_k \mathbb{E}c_{yx|k}/K} \tag{17}$$

and the new MH estimator $\tilde{\Psi}$ to a similar expression $\tilde{\Psi}^{\infty}$ only replacing $c_{xy|k}$ with $\tilde{c}_{xy|k}$.

The reason why $\hat{\Psi}$ is inconsistent is that $\mathbb{E}c_{xy|k} \propto \pi_{x|k}\bar{\pi}_{y|k}$ does NOT hold under arbitrary dependence between items given k, but $\mathbb{E}\tilde{c}_{xyk} \propto \pi_{x|k}\bar{\pi}_{y|k}$ holds, which ensures the consistency of $\tilde{\Psi}$ under the common odds ratio assumption.

Since we introduced additional random effects β_{xk} to account for heterogeneity, we also have to compute $\mathbb{E}_{\boldsymbol{\beta}_k}\pi_{x|k}\bar{\pi}_{y|k}$ in order to obtain an expression for $\mathbb{E}c_{xy|k}$ and $\mathbb{E}\tilde{c}_{xy|k}$.

From $\bar{\pi}_{x|k} = 1 - \pi_{x|k}$ it follows that $\mathrm{Cov}(\pi_{x|k}, \bar{\pi}_{y|k}) = \mathrm{Cov}(\bar{\pi}_{x|k}, \pi_{y|k}) = \mathrm{Cov}(\pi_{x|k}, \pi_{y|k})$. Therefore

$$\mathbb{E}\pi_{x|k}\bar{\pi}_{y|k} = \mathrm{Cov}(\pi_{x|k}, \bar{\pi}_{y|k}) + \mathbb{E}\pi_{x|k}\mathbb{E}\bar{\pi}_{y|k}.$$

Under model (N), we can use the numerical approximation $\mathbb{E}_{\beta_{xk}}\pi_{x|k} \approx \mathrm{expit}\{\delta_{\sigma_x^2}(\alpha_k + \beta_x)\}$ with $\delta_{\sigma^2} = (1+\gamma^2\sigma^2)^{-1/2} \approx (1 + 0.35 \cdot \sigma^2)^{-1/2}$ and $\gamma = 16\sqrt{3}/(15\pi)$ [20] The term $\mathrm{Cov}_{\boldsymbol{\beta}_k}(\pi_{x|k}, \pi_{y|k})$ is zero under model (N), because $\rho = 0$ implies independence between β_{xk} and β_{yk}.

Under model (U),

$$\mathbb{E}_{\beta_{xk}}\pi_{x|k} = \frac{1}{2d}\log\left(\frac{1+\exp(\alpha_k+\beta_x+d)}{1+\exp(\alpha_k+\beta_x-d)}\right)$$

which is approximately $\pi^0_{x|k} := \mathrm{expit}(\alpha_k + \beta_x + \beta_{xk})|_{\beta_{xk}=0}$. Figure 2 shows the quality of this approximation. We only need to be cautious if we apply the above transformation of random effects under model (U) (if $d_1 \neq d_2$ or $\rho \neq 0$) because the difference of two independent uniform distributions $U[-d_1, d_1]$ and $U[-d_2, d_2]$ follows a triangular distribution. However this distribution has less heavy tails than the uniform distribution. Hence we expect that the approximation for such a random effects distribution is even better.

Now we need to compute $\mathrm{Cov}_{\boldsymbol{\beta}_k}(\pi_{x|k}, \pi_{y|k})$ for model (U), as $\rho_{xy} = 0$ does not imply independence between β_{xk} and β_{yk}. We apply a first order Taylor series expansion around $\boldsymbol{\beta}_k = \mathbf{0}$ of function $\mathbf{g}(\boldsymbol{\beta}_k) := (\mathrm{expit}(\alpha_k + \beta_x + \beta_{xk}), \mathrm{expit}(\alpha_k + \beta_y + \beta_{yk}))^T$ and obtain

$$\begin{aligned}\mathrm{Cov}\left(g(\boldsymbol{\beta}_k)\right) &\approx \mathrm{Cov}\left\{\mathbf{g}(\mathbf{0}) + \mathbf{G}(\mathbf{0}) \cdot \boldsymbol{\beta}_\mathbf{k}\right\} \approx \mathbf{G}(\mathbf{0}) \cdot \mathrm{Cov}(\boldsymbol{\beta}_k) \cdot \mathbf{G}(\mathbf{0}) \\ &= \mathbf{G}(\mathbf{0}) \cdot \boldsymbol{\Sigma} \cdot \mathbf{G}(\mathbf{0})\end{aligned}$$

with $\mathbf{G} := \frac{\partial \mathbf{g}(\boldsymbol{\beta}_k)}{\partial \boldsymbol{\beta}_k}$ that yields $\mathbf{G}(\mathbf{0}) = \mathrm{Diag}(\pi^0_{x|k}\bar{\pi}^0_{x|k}, \pi^0_{y|k}\bar{\pi}^0_{y|k})$. It follows

$$\mathrm{Cov}(\pi_{x|k}, \pi_{y|k}) \approx \rho\sigma_x\sigma_y\pi^0_{x|k}\bar{\pi}^0_{x|k}\pi^0_{y|k}\bar{\pi}^0_{y|k}.$$

Hence $\mathrm{Cov}(\pi_{x|k}, \pi_{y|k}) \approx 0$ for $\rho = 0$.

Plugging this into formula (17) yields

$$\Psi^{\infty} \approx \frac{\lim_{K\to\infty}\sum_k \pi^0_{x|k}\bar{\pi}^0_{y|k}/(n_k K)}{\lim_{K\to\infty}\sum_k \pi^0_{y|k}\bar{\pi}^0_{x|k}/(n_k K)} = \exp\{\delta_{\sigma_x^2}(\beta_x - \beta_y)\}$$

because $\exp(\delta_{\sigma_x^2}(\beta_x - \beta_y)) = \frac{\pi^0_{x|k}\bar{\pi}^0_{y|k}}{\pi^0_{y|k}\bar{\pi}^0_{x|k}}$ (a different common odds ratio). We conclude $\log \Psi^\infty = \delta_{\sigma_x^2}(\beta_x - \beta_y) = \delta_{\sigma_x^2} \cdot \log \Psi$, which was stated in Theorem 2 except σ_x^2 has been replaced by $\{\sigma_x^2 + \sigma_y^2 - 2\rho_{xy}\sigma_x\sigma_y\}/2$ due to the transformation.

The numerical approximation for model (N) is based on the result $\mathbb{E}_{\beta_{xk}}\text{expit}(\alpha_x + \beta_{xk} + \beta_x) \approx \text{expit}\{\delta_{\sigma_x^2}(\alpha_x + \beta_x)\}$ derived by Zeger et al. [20]. Under model (U) we use the approximation $\mathbb{E}_{\beta_{xk}}\text{expit}(\alpha_x + \beta_{xk} + \beta_x) \approx \text{expit}(\alpha_x + \beta_x)$. The quality of this approximation can be seen in Figure 2. For $d \leq 1$ the approximation is very good, but for $d \geq 3$ it becomes inaccurate.

This is similar to Figure 2 in [20] where the normal distribution for which the approximation is inaccurate when $\sigma_x^2 \geq 4$. This indicates that a similar approximation might also apply to model (U): $\mathbb{E}_{\beta_{xk}}\text{expit}(\alpha_x + \beta_{xk} + \beta_x) \approx \text{expit}\{\delta_{\sigma_x^2}(\alpha_x + \beta_x)\}$ but with $\delta_{(\cdot)}$ defined differently.

C Appendix

The formula for the generalized covariance estimator is

$$\widehat{\text{Cov}}(\bar{\tilde{L}}_{xy}, \bar{\tilde{L}}_{wz}) = \frac{1}{c^2}\{\tilde{U}^+_{xw} - \tilde{U}^+_{xz} - \tilde{U}^+_{yw} + \tilde{U}^+_{yz}\} \tag{18}$$

with

$$\tilde{U}^+_{xy} = \begin{cases} \tilde{U}^+_{xx} = \tilde{U}_{x++} = \sum_{h,i}\tilde{U}_{xhi} & , x = y \\ \tilde{U}^+_{xy} = \tilde{U}_{+xy} - \tilde{U}_{xy+} - \tilde{U}_{yx+} + \tilde{U}_{xyy} + U^*_{xy} & , x \neq y \end{cases}, \tag{19}$$

where $U^*_{xy} = \sum_{x,h,y,i \text{ distinct}} \tilde{U}_{xhyi}$ for $x \neq y$, otherwise $U^*_{xy} = 0$.

A sketch of this proof is provided here. Subscript "+" denotes summation over that subscript. First we derive

$$\begin{aligned}
&\text{Cov}(\bar{\tilde{L}}_{xy}, \bar{\tilde{L}}_{wz}) = \text{Cov}(1/c\sum_h \tilde{L}_{xh} - \tilde{L}_{yh}, 1/c\sum_i \tilde{L}_{wi} - \tilde{L}_{zi}) \\
&= \frac{1}{c^2}\sum_i \Big\{\text{Cov}(\tilde{L}_{xi}, \tilde{L}_{wi}) + \text{Cov}(\tilde{L}_{yi}, \tilde{L}_{zi}) - \text{Cov}(\tilde{L}_{xi}, \tilde{L}_{zi}) - \text{Cov}(\tilde{L}_{yi}, \tilde{L}_{wi})\Big\} \\
&+ \frac{1}{c^2}\sum_{i \neq h} \Big\{\text{Cov}(\tilde{L}_{xh}, \tilde{L}_{wi}) + \text{Cov}(\tilde{L}_{yh}, \tilde{L}_{zi}) - \text{Cov}(\tilde{L}_{xh}, \tilde{L}_{zi}) - \text{Cov}(\tilde{L}_{yh}, \tilde{L}_{wi})\Big\}
\end{aligned}$$

and express $\sum_{h \neq i} \operatorname{Cov}(\tilde{L}_{xh}, \tilde{L}_{wi})$ as

$$\begin{aligned}
\sum_{h \neq i} \operatorname{Cov}(\tilde{L}_{xh}, \tilde{L}_{wi}) = & \sum_{h} \operatorname{Cov}(\tilde{L}_{xh}, \tilde{L}_{wx}) + \sum_{i} \operatorname{Cov}(\tilde{L}_{xw}, \tilde{L}_{wi}) \\
& (i = x) \qquad\qquad (h = w) \\
& - \operatorname{Cov}(\tilde{L}_{xw}, \tilde{L}_{wx}) + \sum_{\text{distinct indices } x,h,w,i} \operatorname{Cov}(\tilde{L}_{xh}, \tilde{L}_{wi}) \\
& = - \sum_{i} \operatorname{Cov}(\tilde{L}_{xw}, \tilde{L}_{xi}) - \sum_{i} \operatorname{Cov}(\tilde{L}_{wx}, \tilde{L}_{wi}) \\
& + \operatorname{Cov}(\tilde{L}_{xw}, \tilde{L}_{xw}) + \sum_{\text{distinct indices } x,h,w,i} \operatorname{Cov}(\tilde{L}_{xh}, \tilde{L}_{wi})
\end{aligned}$$

These two formulae together provide the basis for equations (18) and (19). For more details of the proof, refer to Suesse [26].

The formulae are identical to Greenland's except formula (19), which contains an additional term U^*_{xy}, because now generally $\operatorname{Cov}(\tilde{L}_{xy}, \tilde{L}_{xy}) \neq 0$. Greenland sampling models a) and b) imply $\operatorname{Cov}(\tilde{L}_{xy}, \tilde{L}_{xy}) = 0$. When comparing Greenland's formula with (19), we see that equation (19) has an additional term S_{xy} for $x \neq y$, because generally $\operatorname{Cov}(\tilde{L}_{xy}, \tilde{L}_{xy}) \neq 0$. We propose to use the bootstrap estimates of covariance $\tilde{U}^*_{xyz}$ and $\tilde{U}^*_{xywz}$ $\tilde{U}_{xyz}$ and $\tilde{U}_{xywz}$, because we are not able to give formulae for the estimators $\tilde{U}_{xyz}$ for $\operatorname{Cov}(\tilde{L}_{xy}, \tilde{L}_{xz})$ and $\tilde{U}_{xywz}$ for $\operatorname{Cov}(\tilde{L}_{xy}, \tilde{L}_{wz})$.